JCBI
Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606
Volume 09 Issue 01
2025

Research Article
https://doi.org/10.56979/901/2025

Unveiling Hidden Themes of Gender Specific Social Anxiety through Linguistic Exploratory Analysis Using LLM and Topic Modeling Techniques

Muhammad Rizwan[1], Saima Noreen Khosa[2*], Maryam Rafiq[2], and Rida Fatima[2]

[1]Institute of Information Technology, Khwaja Fareed University of Engineering and Information Technology, Rahimyarkhan, 64200, Pakistan.
[2]Institute of Computer Science, Khwaja Fareed University of Engineering and Information, Technology, Rahimyarkhan, 64200, Pakistan.
[*]Corresponding Author: Saima Noreen Khosa. Email: saimakhosa@yahoo.com

Received: April 02, 2025 **Accepted:** May 25, 2025

Abstract: This article aims to unveil hidden themes related to social anxiety in gender-specific-manner. For this purpose, dataset of over 12,000 Reddit posts related to social anxiety were used. Traditional preprocessing steps including lemmatization were applied to clean the data. Initially, Llama 3 was employed for zero-shot gender classification using an appropriate prompt to label the posts by gender. The zero-shot classification was then evaluated against human judgment and baseline algorithms. Top2Vec was fine-tuned to identify prevalent linguistic traits and topics within the female and male groups. Various embedding methods were experimented with, and coherence scores were used as evaluation metric for searching best embedding for topic modeling with high coherence score. Doc2vec gives the best coherence score. The optimal settings generated topic vectors with relevant keywords for each gender, highlighting key social anxiety themes. A method was devised to identify the most similar and dissimilar topics for both genders. The analysis revealed significant similarities in male social anxiety posts with female posts in themes of social interaction, mental health, daily activities, dating, and professional communication. Conversely, the least similar topics in female social anxiety posts compared to male posts centered around issues like appearance, facial expressions, school interactions, and strategies for overcoming social anxiety. This analysis underscores the diverse contexts of social anxiety experiences across genders.

Keywords: Social Anxiety; Reddit; Topic Modeling; Top2Vec; Lllama; Zero shot Classification; Gender Classification; Gender Based Social Anxiety

1. Introduction

Social anxiety disorder (SAD) is facing the consistent fear in different social situations like which triggers the emotion of humiliation, rejection and embarrassment [1]. According to Anxiety and Depression Association of (ADAA) [2] approximately 15 million adults in US has to deal with SAD with equal proportion of men and women. Typically emerging around age 13, SAD often goes untreated for extended periods, with a staggering 36% of sufferers waiting 10 years or more before seeking help, according to a 2007 survey by ADAA. These survey and statistics clearly highlight the importance of study SAD with multiple dimensions to provide aid in therapy process. People with Social Anxiety Disorder (SAD) often struggle with forming and maintaining friendships and relationships within their community. This difficulty frequently leads them to isolate themselves, which can, in turn, trigger additional mental health disorders such as depression or generalized anxiety. SAD, a form of social phobia, profoundly affects individuals' ability to interact in social settings[3], [4] While extensive research has been conducted on various dimensions of social anxiety, one significant area that remains under explored is gender-based social anxiety as observed through social media data. While previous research

has explored gender-based social anxiety [5], [6], [7], to the best of our knowledge, no studies have utilized social media data or Reddit with gender-specific labeling. So utilizing the zero-shot capability of LLM such as Lllama3, this study also produced gender labeled Reddit data belongs to social anxiety. Understanding how social anxiety manifests differently in men and women can provide valuable insights into tailored treatment approaches. This study aims to address this gap by analyzing social media data to identify and differentiate gender-specific traits associated with social anxiety. This study used the state-of-the-art NLP tools and techniques such as Lllama3 and topic modeling. The findings will assist clinicians in developing more effective, personalized treatment plans for individuals based on their gender-specific social anxiety characteristics. The following contributions have been made in this article:

- Performed gender-based labeling of social anxiety Reddit data using Large Language Models (LLMs) and further analyze its performance using human judgement and baseline algorithms.
- Performed topic modeling to identify gender specific topics and themes through fine-tuning of Top2Vec using multiple embedding techniques and coherence score as performing indicator.
- After fine tuning topic modeling and calculation of topic vectors for both male and females, word cloud of top topics generated for male and female social anxiety.
- Devised a method to identify the most similar and dissimilar topic vectors for males compared to females and vice versa.
- The rest of the section organized as related work, method with result and analysis of experimentation,
- Conclusion and lastly limitation and future work.

2. Related Work

Social media platforms, particularly Reddit, have emerged as significant resources for exploring various mental health issues, including depression, anxiety, and social anxiety. These platforms offer rich, user generated content that searchers are increasingly leveraging to understand mental health dynamics more deeply. As social anxiety is a common issue impacting personal and professional aspects of life, understanding its physical and emotional symptoms [8]

is crucial for accurate diagnosis and effective treatment. The study[9] examines social anxiety symptoms using data from the Mayo Clinic and a large Reddit dataset. By employing BART-based multi-label zero-shot classification, the research identifies and measures symptom prevalence, revealing ”Trembling” as a frequent physical symptom and ”Fear of being judged negatively” as a common emotional symptom. These findings provide valuable insights into the nature of social anxiety, aiding in the development of targeted clinical interventions.

Another study[10] explores how text message language patterns can reflect mental health conditions. By analyzing text messages from 335 adults over 16 weeks using tools such as LIWC and the NRC Emotion Lexicon, the study identified significant correlations between language features and mental health symptoms. The findings reveal that depressive symptoms are linked to negative language concerning anticipation, trust, social processes, and affiliation, while generalized anxiety shows positive associations with these features. Social anxiety, on the other hand, is uniquely related to expressions of anger, sexual language, and swearing. These insights underscore the potential of text message analysis to uncover cognitive-behavioral patterns and inform digital mental health interventions. Other examples of similar studies are [11], [12].

One notable advancement in this field is presented in the study [13] Depression, with its profound effects on both physical and mental health, has been a major focus for automated detection efforts. Traditional methods in depression detection often rely on basic emotional analysis but may miss nuanced high-level emotional semantics. This study introduces an innovative emotion-based attention network, combining a semantic under-standing network with an emotion understanding network [14]. This approach is designed to capture complex emotional subtleties and significantly enhance the accuracy of depression detection. Evaluated on a Reddit dataset, the proposed model demonstrated exceptional performance, achieving an accuracy of 91.30%, a precision of 91.91%, and a recall of 96.15%. These results highlight the effectiveness of integrating advanced emotional semantic information for more reliable depression detection. Other examples of similar studies are[15], [16].

The impact of social media use on mental health is further explored in the study[17]. This research examines how problematic social media use affects real-life social support and mental health. Through an online survey, the study found that problematic social media use correlates with diminished real-life support and increased reliance on social media for support. Importantly, real-life support was associated with lower levels of depression, anxiety, and social isolation, whereas social media support did not show the same protective effects. These findings highlight the critical role of face-to-face social support [18] in mitigating the adverse mental health impacts of problematic social media use and suggest areas for further research and intervention. A systematic review titled "Social Media Use, Social Anxiety, and Loneliness: A Systematic Review"[19] investigates the relationship between social media use and mental health, focusing on social anxiety and loneliness. The review indicates that individuals who are socially anxious and lonely often engage with social media in problematic ways, seeking online interactions to compensate for insufficient in-person connections. The review also notes that most existing research relies on self-reported and cross-sectional data, calling for more experimental and longitudinal studies to better understand the bidirectional relationships between social media use, social anxiety, and loneliness. The influence of social media on social anxiety was also examined in the study[20]. This research used Latent Dirichlet Allocation (LDA) to analyze topics and emotions in Reddit posts about social anxiety before and during the COVID-19 pandemic. The study identified thirteen key topics, including social interactions and coping mechanisms, and found that the primary emotions expressed were anticipation, trust, and fear. Notably, the sentiment of posts remained relatively consistent across the two periods, though some topics saw a shift in the nature of comments. These findings highlight the stability of online discussions about social anxiety and suggest that further research could explore the pandemic's specific impact on these discussions and the potential benefits of online community engagement for those with social anxiety. [21] investigates how personality traits correlate with internet and social media addiction. The study compared symptoms of internet and social media addiction with alexithymia, narcissism, and social anxiety among 217 young adults. The results indicated that while social anxiety and narcissism predicted both internet and social media addiction symptoms, alexithymia was associated only with internet addiction. This distinction suggests that internet addiction may encompass broader issues compared to social media-specific behaviors.

The study [22] focuses on gender-specific expressions of mental health symptoms related to cardiovascular disease on Reddit. By using a knowledge assisted RoBERTa-based bi-encoder model, the study aimed to enhance the accuracy of identifying gender specific language in mental health symptoms. The model demonstrated high performance in predicting mental health issue labels and gender labels, improving recall rates for both, thus advancing the precision of mental health symptom detection on social media. In another study[23], they utilized NLP to analyze posts from mental health support groups on Reddit during the COVID-19 pandemic. The study revealed a rise in posts related to economic stress and isolation, with notable increases in suicidality and loneliness. The findings underscore the effectiveness of NLP in identifying at-risk users and emerging concerns, which can aid in resource allocation and support during pandemics and other major events. Together, these studies illustrate the rich potential of social media platforms like Reddit for advancing our understanding of mental health issues in general and specifically social anxiety.

3. Methods

Figure 1 shows the detailed workflow and methodology of this study. The rest of the subsection elaborate all steps mentioned in the workflow figure.

3.1. Dataset

The dataset used in this research, as described by [23], was obtained using Reddit's Pushshift API. This dataset consist of posts from 15 distinct subreddits focused on various mental health communities, covering a wide range of mental health issues. For this study's specific objectives, we focused on the subreddit r/ social anxiety, concentrating on discussions related to social anxiety disorder. The dataset consists of 12,277 text documents from the r/socialanxiety subreddit, capturing content related to social anxiety posted between 2018 and 2019 on Reddit platform. Within this dataset, individuals share their real-life experiences, opinions, and symptoms related to social anxiety. Each document is associated with

a unique user, resulting in a dataset diversified with contributions from more than 12,000 unique users, enhancing reliability of the results at the end. Our purpose in utilizing this dataset is to explore the unique perspectives and challenges expressed by individuals within these online communities. The central objective is to analyze gender-specific traits of social anxiety within the r/social anxiety subreddit. How ever, the dataset does not identify which posts belong to males or females.

To address this, we used zero-shot classification by the LLaMA 3 large language model to classify the data into two categories: male text and female text. Next section elaborate this process in detail.

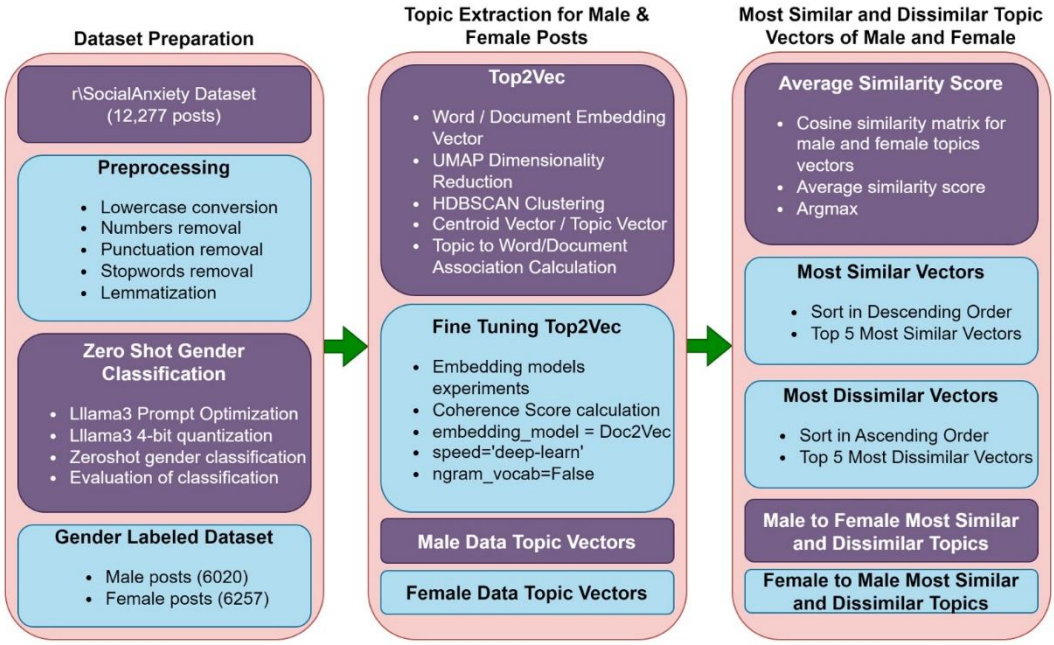

Figure 1. Detailed workflow of this study.

3.2. Lllama3 Zero Shot Gender Classification

Lllama3 is used for zero shot classification of gender within the dataset for all social anxiety posts. Llama3 [24], developed by Meta AI, is the advanced successor to LLaMA 2, incorporating numerous improvements while maintaining the foundational transformer architecture of its predecessor. LLaMA 3 is designed to offer enhanced model efficiency and accuracy, and it has been trained on a significantly larger dataset. This extensive training allows the model to better understand context and handle complex queries more effectively.

When compared to other open-source language models such as Mistral-7B and Gemma-7B, LLaMA 3 shows superior performance. The model is available in two versions: one with 8 billion parameters and another with 70 billion parameters. For our study, we utilized the 4-bit quantized version of the 8 billion parameter model, officially listed as 'meta llama/Meta-Llama-3-8B-Instruct'on Hugging Face.

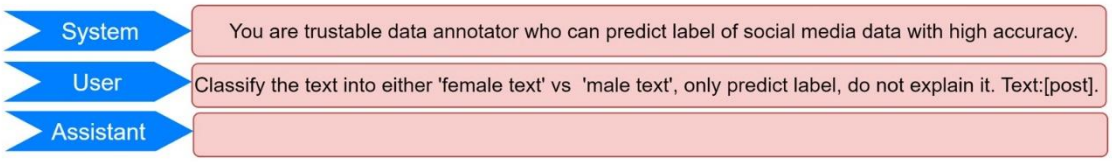

Figure 2. Prompt Template Used with LLaMA 3 Model for Gender Classification Task

To determine the most effective prompt for zero shot gender classification, we experimented with several prompts by manually observing the labels (male/female) generated for randomly selected social anxiety-related posts. The final prompt, which yielded the best results for our classification task can be seen in Figure 2.

For zero shot classification experiments, we utilized LLaMA 3 with the Hugging Face text generation pipeline, configured with the parameters as max new tokens=15, do sample=True, return full text=False, temperature=0.7, top k=50, top p=0.95. This configuration allowed us to generate concise and varied text

samples from a given prompt. To assess the Lllama3 zero-shot gender classification performance, we first manually observed random post content and their predicted labels. For instance, if you look at the sample posts classified by Llama3 into male or female in Table 1, none of the post content explicitly mentioned the gender, but Llama3 deduced the gender from the context quite well. For example, in post 1, the person mentions 'not having a girlfriend', which Llama3 correctly deduced as male. Similarly, in post3, the phrase "love to find yourself a girl" indicates male post, which Llama3 correctly identifies. In post 4, Llama3 accurately identified the gender as female because the post mentioned "husband's friend," implying it was a female post. However, post 2 did not provide any apparent clues about the gender, but Llama3 classified it as a female post as well. After labeling the data with zero-shot gender classification, we further validated Llama3's performance by applying Naive Bayes and Logistic Regression, using 80 percent of the data for training and 20 percent for testing. Naive Bayes achieved 62 percent accuracy, and Logistic Regression achieved 64 percent accuracy. These results are quite decent as these algorithms serve as baselines, which implicate that Lllama3 performed really well.

3.3. Preprocessing and Cleaning

In this study, we utilized the spaCy library [25] to preprocess text data. First, we loaded a pre-trained Englis language model from spaCy (en_core _web_sm) to facilitate tasks such as tokenization, part-of-speech tagging and lemmatization. Next, we defined a custom function to clean and lemmatize individual text posts. This process included removing extra white spaces, stop words, digits, and punctuation, as well as converting all text to lower case. Unlike stemming, which reduces words to their root form, lemmatization was used to convert words to their base or dictionary form, preserving meaningful variations. Finally, we applied this cleaning and lemmatization function to the 'post column of our dataset. This resulted in a new column containing standardized text data, ready for further analysis. By implementing these preprocessing steps, we ensured that the text data was normalized, thereby improving the accuracy and reliability of our subsequent linguistic analysis and topic modeling.

3.4. Topic Modeling for Gender Specific Insights

As the main purpose of this study is to identify the gender-specific traits of social anxiety, we first aimed to extract prominent topics or themes within each dataset group, namely 'male post' and 'female post', related to social anxiety. Then we comparatively analyze the similarities and dissimilarities of male and female in social anxiety. For the topic modeling task, we used and fine-tuned Top2Vec to achieve better topic representation in each group. Top2Vec is a state-of-the-art topic modeling algorithm that identifies topics within a text corpus using document embeddings.

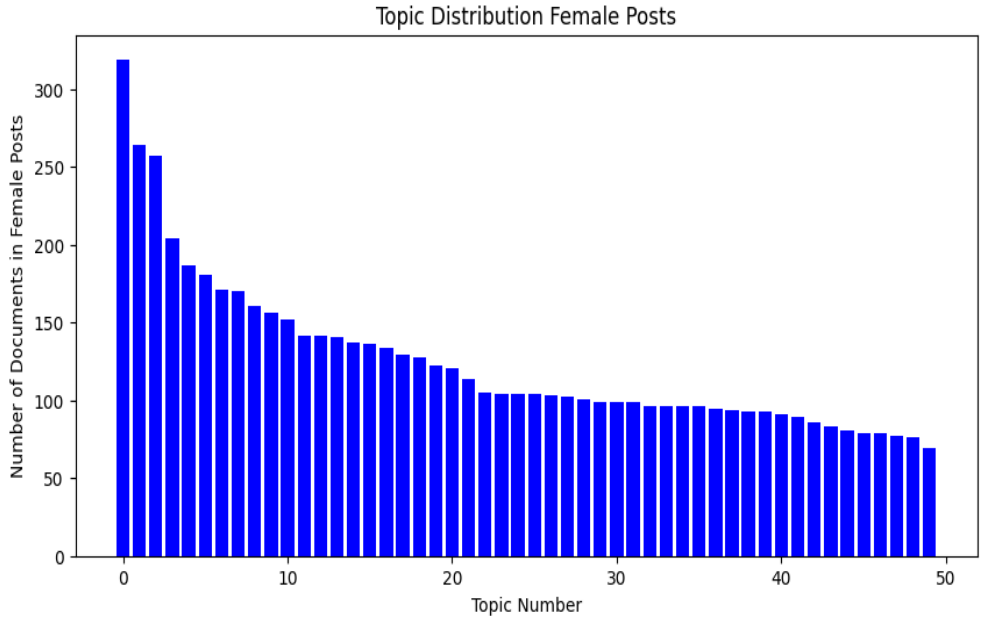

Figure 3. Topic Distribution of Female Posts

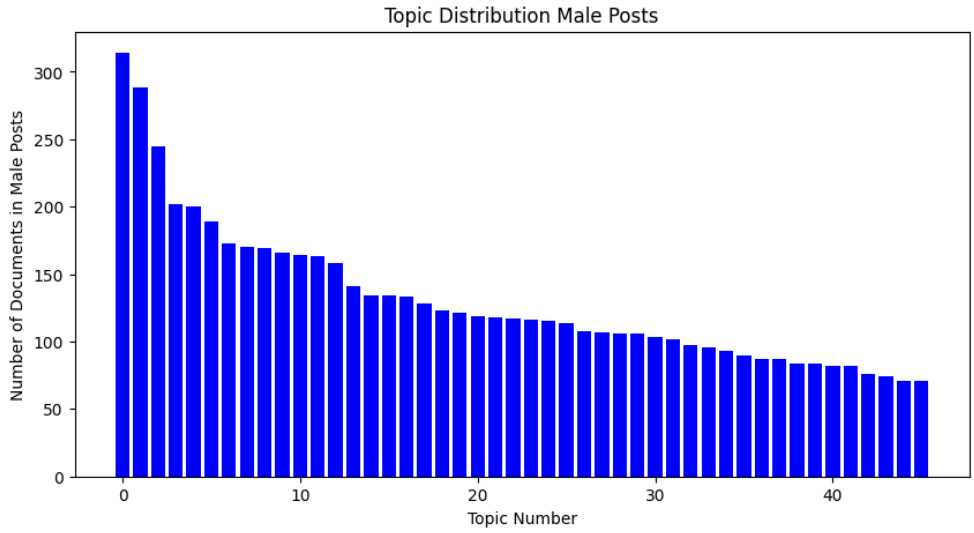

Figure 1. Topic Distribution of Male Posts

With the preprocessed data ready as described in the previous section, the first step of topic modeling using Top2Vec is to compute the embedding vector of each document to convert it into a numerical representation. we optimized the Top2Vec algorithm for topic modeling by selecting specific parameters. Using speed="deep- learn" allowed me to generate high-quality, context aware embeddings enhancing semantic understanding. Choosing ngram vocab=False simplified the vocabulary to individual words, focusing the analysis on the most relevant terms while reducing computational complexity. Setting workers=32 enabled parallel processing, significantly speeding up computations and optimizing resource utilization. For embedding purposes, Top2Vec supports various BERT like language models and Doc2Vec, among others. Finally, by setting embedding model='doc2vec' provided effective document representations, preserving context and ensuring accurate theme identification.

Table 1. Posts and their corresponding gender label where 'M' represent male and 'F' represent female

Sr.#	Post	Gender
1	No social media. I have no friends or social life, much less a girlfriend. Because of this I don't have any social media, for I wouldn't have anyone to add or anything to post. But when I see everyone else on their phones all the time, I keep wondering what is there to Instagram, etc that's so interesting? What are people up to?	M
2	Doesn't it suck when you are talking to someone (but not really because of social anxiety) and the school day ends and you are sitting by your actual friend(s) and they only say goodbye to your friend(s) and not you? Happens a lot. A lot.	F
3	What to do when feeling anxious? Hey people of Reddit, do you have any tips for what to do when you are feeling extremely anxious and don't want to do anything but would love to just find yourself a girl and be happy?	M
4	In public with my husband's friends, so anxious. Suggestions for how I'm normal? Lol I'm in public with my husband's friends, I'm drinking, telling stories, getting embarrassed, trying to"be cool", I feel stupid. I don't think I'm doing this right. Writing for a moment to look at my phone and realize by your responses I'm ok.	F

This is a crucial step, as embeddings provide the semantic context of the documents. In this study, we evaluated the coherence scores of various embedding methods employed in the Top2Vec model to determine their effectiveness in capturing meaningful topic representations. The embeddings experimented include the Universal Sentence Encoder, Universal Sentence Encoder Multilingual,

DistilUSE Base Multilingual Cased, and Doc2Vec. Each embedding method was assessed based on its coherence score, a metric that reflects the consistency and inter pretability of the generated topics. As shown in Table 2, the coherence scores for these embeddings were 0.28, 0.21, 0.27, and 0.36, respectively. Notably, Doc2Vec achieved the highest coherence score of 0.36, indicating its superior performance in producing coherent and interpretable topics compared to the other embedding methods evaluated.

Table 2. :Embedding methods in Top2Vec and their coherence scores

Embedding	Coherence Score
Universal Sentence Encoder	0.28
Universal Sentence Encoder Multilingual	0.21
DistilUSE Base Multilingual Case	0.27
Doc2Vec	**0.36**

The next step was to perform clustering on these embedding vectors. However, high-dimensional vectors need to be converted into a lower-dimensional space to make the process computationally efficient. For dimensionality reduction, Top2Vec uses UMAP (Uniform Manifold Approximation and Projection), which effectively reduces the dimensionality while preserving the topological structure of the data. Following dimensionality reduction, Top2Vec employs HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) to identify dense regions in the low-dimensional vector space and form clusters. For each cluster identified by HDBSCAN, the centroid vector is computed, representing the central theme of the topic associated with the documents within the cluster. This centroid vector is known as the topic vector. To make the topics human-understandable,

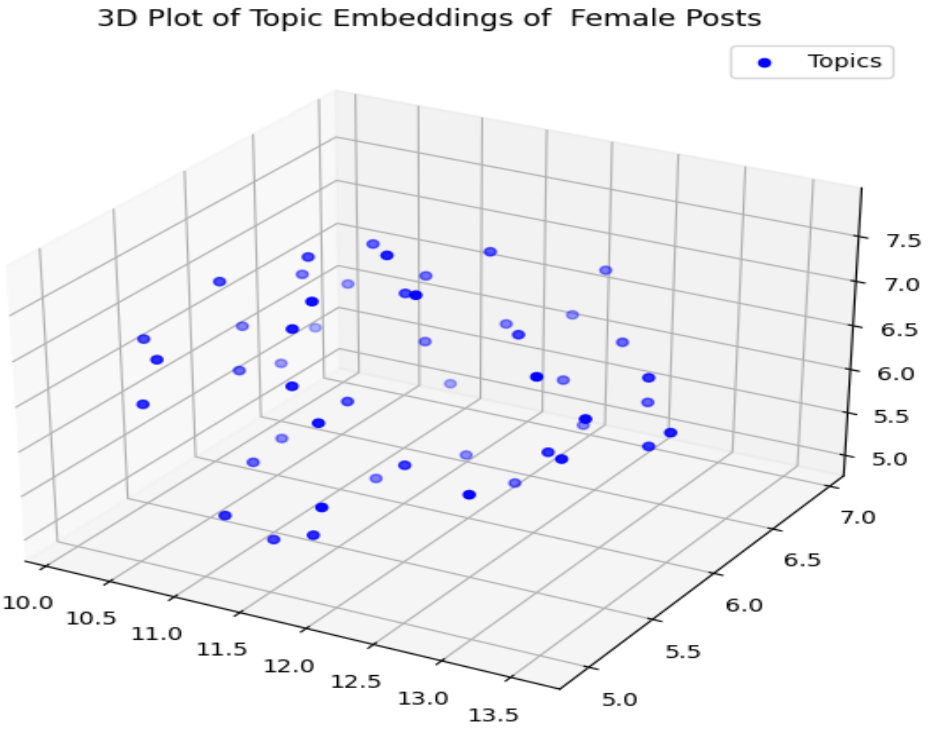

Figure 5. 3D Plot of Topic Embeddings of Female Posts

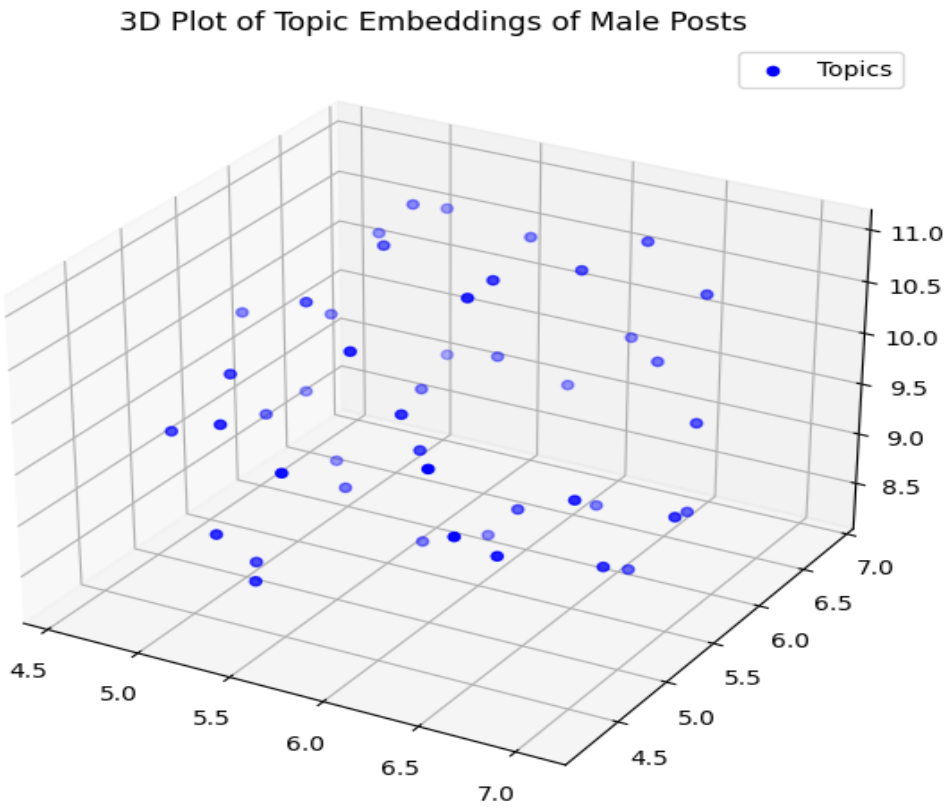

Figure 2. 3D Plot of Topic Embeddings of Male Posts

The closest words to each topic vector are identified using cosine similarity. In the final step, the association of each document with the identified topics is calculated based on the words associated with each topic. One document can be associated with multiple topics in this approach. By following the above procedure, we computed the topics in both the female and male posts. Figure 5 and 6 shows the 3-dimensional plot of topics embedding in female and male posts respectively. The plots shows that the topic are well distributed and does not overlap, thus generated quite good topics.

3.4.1. Discussion about results

In this section the comparative analysis of top 10 topics generated in both male and female is done. Figure 4 shows the distribution to topics among female posts; we consider top 10 topics which are discussed. If you look at the top 10 topics generated in the case of social anxiety in females, the first topic is about work-related anxiety issues such as fear of interviews, evaluations, interaction with new colleagues, and promotions. The second topic is about therapy and medication for anxiety and consultation with a psychologist. The third topic is about fears in social interactions at parties as well as difficulty in communication and conversation. The 4th topic showcases the fear of being judged and being the center of attention in public speaking, such as presentations, whereas the fifth topic is about online communication, messaging, feelings of being misunderstood, and difficulty initiating online conversations. The sixth topic is about facing physical symptoms related to social anxiety, such as trembling, sweating, and blushing, as well as overthinking, whereas the 7th topic is related to being alone and fear of new people and places. Topic 8th is about difficulty in communication and expressing oneself with fear of being rejected and ignored. Topic 9th is about difficulty in self-expression and being introvert whereas 10th topic is about fear of being perceived as awkward during eye contact. When analyzing the top 10 topics generated in posts by males related to social anxiety, all other topics appear in female posts as well, except for two. The fourth topic highlights anxiety experienced in social settings involving alcohol and parties. Keywords such as ”drunk,” ”party,” ”bar,” ”invite,” and ”fun” suggest the discomfort and fear of socializing in such environments, often leading to avoidance behaviors. This indicates that drunk issues are more related to males compared to females. The fifth topic exhibits severe anxiety and depression, often linked to childhood trauma and family issues. Keywords like ”suicidal,”

"child," "parent," "depression," and "abuse" reveal the deep-rooted emotional struggles that contribute to anxiety. But for systematic comparative analysis of female vs male we opted the approach described in the next section.

3.5. Methods to Calculate Most Similar and Dissimilar

Topics among Males and Females In the previous step, we obtained the male topic vectors array and the female topic vectors array where male topic vectors is a 2D NumPy array of shape M ×D, where M is the number of topic vectors

in the male posts set, and D is the dimensionality of each vector, similarly female topic vectors is a 2D NumPy array of shape F ×D, where F is the number of topic vectors in the female posts set, and D is the dimensionality of each vector.

For each vector v_i in male topic vectors and each vector u_j in female topic vectors, we computed the cosine similarity $sim(v_i, u_j)$ as:

$$\text{sim(vi, u j)}=\frac{V_i \quad U_j}{\|V_i\| \quad \|U_j\|} \tag{1}$$

Where $v_i \cdot u_j$ is the dot product of vectors v_i and u_j, and $\|v_i\|$and $\|u_j\|$ are their Euclidean norms. Then we computed cosine similarity matrix S of shape M×F, where each element s_{ij} represents the cosine similarity between the i-th vector in male topic vectors and the j-th vector in female topic vectors. Each value in this matrix represents the cosine similarity between pairs of topic vectors, where values close to 1 indicate high similarity and values close to 1 indicate high dissimilarity. Figure 7 shows the heatmap of this cosine similarity matrix.

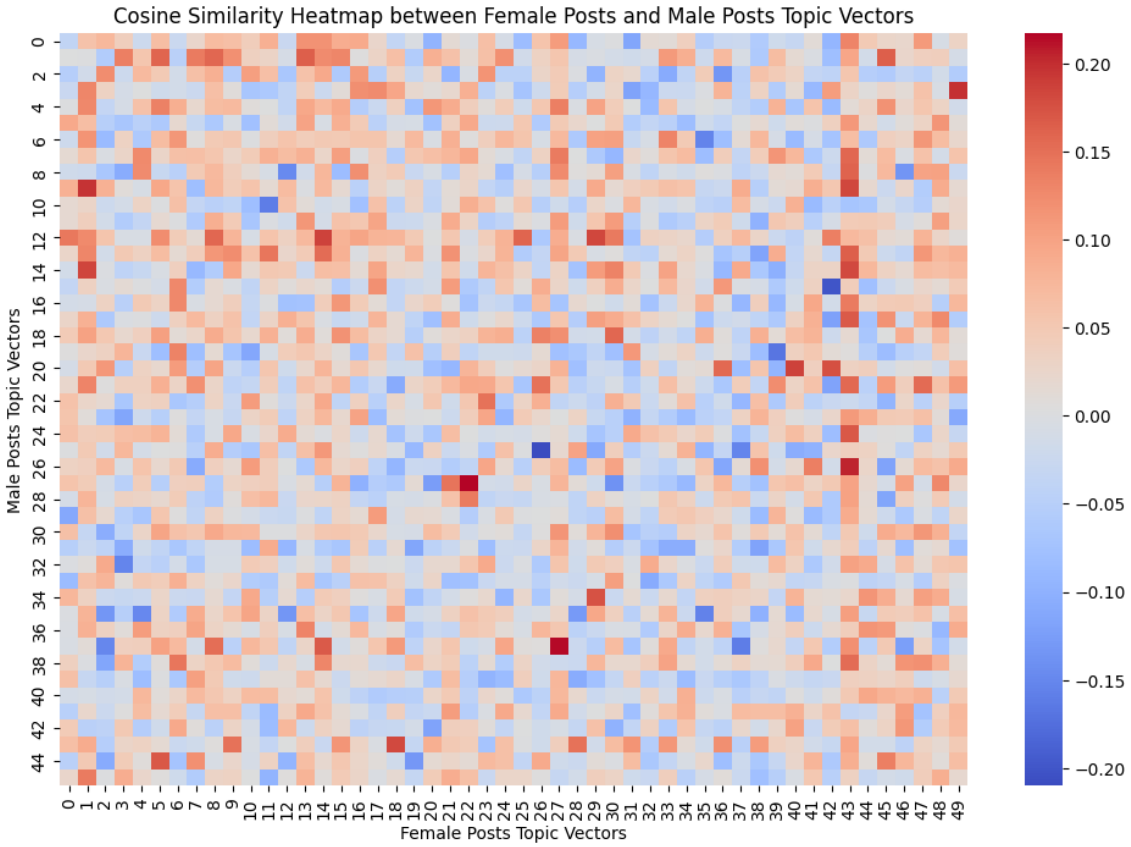

Figure 3. Similarity Heatmap of Male vs Female posts in Social Anxiety

$$S = \begin{vmatrix} s_{11} & s_{12} & \dots \dots s_{1F} \\ s_{21} & s_{22} & \dots \dots s_{2F} \\ s_{M1} & s_{M2} & \dots \dots S_{MF} \end{vmatrix} \tag{2}$$

In next step for each vector v_i in male topic vectors, we calculated the average similarity score sim_i with all vectors in female topic vectors:

$$sim_i = \frac{1}{f} \sum_{j=1}^{f} \binom{n}{k} s_{ij} \quad (3)$$

Lastly, we sorted the indices of male topic vectors based on their average similarity scores sim_i in descending order [26] to identify the most similar topic vectors and sorted the indices of male topic vectors based on their average similarity scores sim_i in ascending order to identify the most dissimilar topic vectors. Table ?? shows the keywords belongs to topmost similar and dissimilar male topics to female posts based on their average similarity as described above. Similarly same procedure adopted to calculate the most similar and dissimilar topic in female post to male post based on their average similarity score described above. Table ?? shows the keywords belongs to topmost similar and dissimilar female topics to male posts. This analysis gives interesting comparative insights of gender based social anxiety.

3.6. Discussion about Results

Finally, from both the descending and ascending lists of topic vectors, we selected the top 5 vectors from each list. The ascending list, which ranks vectors by increasing distance, provided the 5 most similar topic vectors. These vectors represent the maximum similarity in social anxiety traits between males and females from male side, indicating common themes or characteristics shared across genders. Conversely, the descending list, which ranks vectors by decreasing distance, revealed the 5 most dissimilar topic vectors. These vectors highlight the minimum similarity in social anxiety traits between males and females from male side, pointing to distinct, gender-specific themes.

By analyzing these top vectors, we gained insights into both the shared and unique aspects of social anxiety across genders. When we analyzed male social anxiety posts, we identified the topics most similar to female posts based on average similarity scores. The top-ranked topic, with a score of 0.0519, includes keywords such as 'naturally', 'personality', 'boring', 'tend', 'connection', and 'socialize' shows themes around social interaction. The second-ranked topic, scoring 0.0414, includes terms like 'medication', 'prescribe', 'therapy', 'doctor', and 'anxiety', emphasizing medical treatment and mental health. Another topic, with a score of 0.0367, features words such as 'shopping', 'grocery', 'store', and 'cashier', focusing on daily activities and public interactions.

Similarly when we examined the most similar topics in female social anxiety posts compared to male posts, several key themes emerged based on average similarity scores. The highest similarity score of 0.0808 was observed for topics involving general social media interactions, with keywords like 'video', 'watch', 'game', 'reddit', and 'emotion'. The second-ranked topic, with an average similarity score of 0.0526, centered on mental health and treatment, featuring terms like 'medication', 'therapy', 'psychiatrist', and 'doctor'. The third topic, with a similarity score of 0.0483, pertained to social interactions and relationships in high school, with keywords like 'popular', 'friend', 'conversation', and 'shyness'. The fourth topic, with a similarity score of 0.0419, highlighted personality and social behaviors, including 'extrovert', 'introvert', 'habit', and 'behavior'.

Table 2. Top 5 Male Topics with highest average similarity scores with Females Topics

Rank	Vector	Similarity	Keywords
1	12	0.0519	naturally, personality, boring, tend, connection, socialize, bore, question, task, ability, communication, discussion, fairly, difficult, talking, impression, sport, frustrating, small, opinion, dislike, communicate, behavior, diagnose, drain, topic, connect, acquaintance, other, context, silence, fake, conversation, example, mood, decision, introvert, certain, circle, clear, stressful, negative, overthink, voice, seek, overcome, art, relate, brain, participate
2	1	0.0414	mg, medication, prescribe, effect, drug, psychiatrist, therapy, doctor, med, severe, diagnose, symptom, disorder, cure, cbd, pill, reduce, depression, daily, research, therapist, appointment, treatment, sa, health,

			side, abuse, psychologist, session, take, anxiety, suggest, helpful, suggestion, benefit, issue, exposure, overcome, mental, solution, difference, half, recommend, cope, help, tough, seek, illness, personally, mood
3	18	0.0367	card, shopping, grocery, store, shop, buy, cashier, pay, line, park, yesterday, grab, wait, lady, door, cancel, clothe, car, walk, order, coffee, behind, today, employee, bus, realise, money, busy, early, customer, must, ok, food, retail, forgot, rush, wife, late, embarrassed, left, neighbor, slow, street, minute, window, decide, town, freaking, teenager, near
4	7	0.0344	tinder, virgin, date, kiss, match, girl, attractive, sex, relationship, attract, app, girlfriend, male, ex, female, gf, ugly, meet, woman, interested, texte, decent, hot, message, approach, interest, confidence, email, cute, boyfriend, insecure, talking, scare, cbd, medium, pretty, drunk, texting, account, snapchat, man, text, kinda, reject, nerve, common, send, rate, idk, scared
5	13	0.0340	regard, action, communication, destroy, speech, advance, general, interact, professional, overcome, miserable, lack, degree, difficult, may, fire, depression, tough, decision, average, severe, condition, professor, manage, serious, raise, negative, society, massive, disorder, colleague, child, specific, discussion, psychologist, field, doctor, subreddit, crippling, amp, major, emotional, wedding, program, illness, tend, relate, level, connection, helpful

Table 3. Top 5 Male Topics with lowest average similarity scores with Females Topics

Rank	Vector	Similarity	Keywords
1	31	-0.0287	haircut, hair, cut, sweat, embarrassed, shop, appearance, proud, appointment, professional, customer, money, silly, average, embarrassing, nowhere, black, short, clothe, lady, pay, fat, ugh, natural, conscious, buy, pick, min, minute, father, realise, luck, discord, fool, wear, all, dread, yo, pathetic, breath, shaky, seat, straight, drain, local, forgot, race, look, rate, ridiculous
2	39	-0.0103	facial, expression, irrational, face, control, tense, acquaintance, nervousness, blush, uncomfortable, nervous, engage, body, overthinke, assume, smile, symptom, exposure, slightly, mouth, solution, eye, true, maintain, extrovert, stare, everytime, condition, personality, daily, weird, conscious, trying, breathe, shame, look, stutter, fear, appearance, angry, contact, creep, relationship, impression, highschool, mg, might,

			insecure, hot, describe
3	19	-0.0102	creep, cute, girl, crush, approach, texte, creepy, ugly, female, beautiful, attractive, contact, Facebook, attract, number, weirdo, chance, she, reject, classmate, male, interest, message, date, virgin, boy, tinder, sex, courage, loser, relationship, send, snapchat, dude, ignore, stare, grade, app, bore, guy, straight, interested, shy, girlfriend, talk, Instagram, picture, pretty, kinda, catch
4	10	-0.0078	senior, grade, th, junior, school, highschool, freshman, high, bully, elementary, middle, college, friend, year, graduate, class, popular, group, semester, acquaintance, extrovert, lunch, classmate, apart, introvert, parent, alot, blame, student, hang, summer, somewhat, new, nearly, girl, invite, gain, since, boy, grow, outgoing, period, move, ton, friendship, crush, join, band, loser, usual
5	25	-0.0048	zone, comfort, step, opposite, popular, title, overcome, psychologist, trick, along, push, female, environment, hello, YouTube, confidence, childhood, outside, conversation, direction, relax, build, alot, strange, alright, club, ease, suggest, bar, ball, rid, remind, interest, massive, success, topic, dead, breathe, enjoy, solution, action, condition, flow, male, free, safe, tip, effort, slightly, curious

Table 4. Top 5 Female Topics with highest average similarity scores with Females Topics

Rank	Vector	Similarity	Keywords
1	43	0.0808	video, watch, game, reddit, emotion, play, internet, comfort, create, music, post, study, medium, meme, movie, share, app, goal, recommend, rid, haircut, story, topic, society, loneliness, write, interest, facebook, subreddit, focus, boring, feeling, random, compare, hello, instagram, escape, symptom, easy, hope, depressed, side, somebody, ton, shell, delete, normally, curious, note, quick
2	1	0.0526	medication, prescribe, therapist, therapy, psychiatrist, med, cbt, treatment, doctor, effect, disorder, appointment, recommend, seek, diagnose, session, health, symptom, drug, severe, curious, suffer, suggest, physical, psychologist, weight, depression, side, greatly, progress, mental, hello, success, helpful, anxiety, help, experence, affect, list, option, professional, solution,

			book, specifically, difference, mood, finally, cope, wonder, mainly
3	27	0.0483	popular, carry, kid, ppl, facebook, quiet, highschool, overthink, friend, random, conversation, outgoing, boring, funny, common, everybody, crush, shyness, dislike, relate, interesting, contribute, silence, lonely, talking, text, fit, interest, alot, comment, instagram, personality, boy, somehow, annoy, friendship, cool, definitely, texte, setting, group, sub, chat, joke, talk, medium, acquaintance, stuff, picture, girl
4	14	0.0419	extrovert, introvert, extroverte, habit, exam, circle, large, space, shy, child, naturally, create, service, personality, outgoing, grow, particular, contribute, throughout, behavior, apart, study, meme, term, fairly, possible, customer, highschool, talkative, natural, world, laugh, center, react, alcohol, phobia, reality, overthink, fit, society, socialize, exist, prefer, lack, classroom, perceive, realize, follow, result, greet
5	13	0.0403	psychologist, sentence, angry, psychiatrist, crippling, welcome, seriously, breath, they, test, presentation, require, symptom, voice, phobia, haircut, fairly, research, song, compliment, frustrating, severe, disorder, check, burden, apparently, quickly, member, diagnose, drug, word, answer, slow, twice, busy, catch, effect, progress, participate, language, upset, speech, prepare, customer, blank, you, sub, insecurity, english, annoy

Table 5. Top 5 Female Topics with lowest average similarity scores with Males Topics

Rank	Vector	Similarity	Keywords
1	28	-0.0200	concert, band, stressful, drive, field, dance, song, music, haircut, movie, trip, show, race, four, send, overwhelmed, picture, week, husband, money, instagram, cool, tomorrow, enter, excuse, last, car, terrified, apartment, email, greet, touch, drunk, early, psychologist, late, texte, meeting, freak, play, text, chat, ugh, tldr, service, interest, watch, volunteer, list, appointment
2	32	-0.0162	roommate, apartment, move, room, live, bathroom, college, semester, city, freshman, country, hangout, enter, uni, campus, house, boyfriend, partner, stay, impression, lunch, home, alone, street, program, terrified, chat, door, leave, night, town, outgoing, nerve, invite, university, total, within, grocery, hi, wake, absolutely, watch, car, four, food, student, month, interesting, eat, drive

3	37	-0.0158	holiday, family, town, dinner, art, house, member, visit, alot, dress, activity, quick, circle, weekend, big, gym, background, top, rarely, mother, party, dread, boyfriend, adult, extra, kick, band, sub, brother, going, account, acquaintance, invite, relax, bf, crowd, trip, event, country, awkwardness, wear, hair, skip, prefer, guess, grocery, hide, happy, greet, besides
4	42	-0.0153	buy, store, cashier, shop, food, order, clothe, grocery, alcohol, walk, retail, step, street, wear, decide, proud, worried, car, success, anywhere, line, gym, simple, campus, minute, today, drive, apartment, lady, service, coffee, reality, judge, dress, restaurant, customer, small, extra, conscious, hair, crowd, half, dance, dog, ridiculous, appointment, anybody, outside, random, difference
5	35	-0.0130	cute, ride, crush, girl, boy, instagram, guy, blush, pick, drive, he, next, car, number, she, proud, ask, bus, awkward, talking, age, sign, manager, account, shop, super, minute, yesterday, weirdo, around, zero, attractive, straight, texte, red, joke, common, couple, teach, date,yes, smile, hi, coworker, picture, yell, cool, class, generally, introduce

Lastly, the fifth topic, with a similarity score of 0.0403, focused on psychological and physical symptoms, with keywords such as 'psychologist', 'angry', 'crippling', and 'symptom'. These results emphasize the unique contexts and situations that contribute to anxiety in female social anxiety posts.

3.6.1. Distinct Gender Based Patterns

Analyzing the most dissimilar topics for both females and males mentioned in Table 6 and Table 4 reveals a distinct pattern. The topics and keywords in female social anxiety posts that are most dissimilar to male posts tend emotional experiences. In contrast, male social anxiety posts that are most dissimilar to female posts tend to focus on appearance, social interactions, and feelings of inadequacy. Female posts often mention specific social anxiety in situations like concerts, trips, and meetings, while male posts exhibit social anxiety regarding personal characteristics like haircuts, facial expressions, and body language. Female posts also discuss emotional experiences like feeling overwhelmed, stressed, and anxious in multiple public situations, whereas male posts exhibits feelings of embarrassment, nervousness and self-consciousness in public situations.

4. Conclusion

In this study our aim is to unveil hidden themes of gender-specific social anxiety by analyzing linguistic traits using LLM and topic modeling techniques using the Reddit social anxiety data. The results highlight distinct patterns of social anxiety in males and females, offering valuable insights how this affects in a gender diverse way. The implications of these findings are useful for clinical practice. Understanding the gender-specific pattern of social anxiety can be used in more targeted interventions. For instance, therapeutic approaches can be tailored to address specific fears and challenges identified in each gender, potentially leading to better outcomes for individuals suffering from social anxiety.

This study also demonstrates the potential of advanced analytical techniques like LLMs and topic modeling to uncover complex patterns in mental health conditions using textual social media data. By shedding light on the gender-specific themes of social anxiety, this research contributes to a deeper

understanding of the gender specific social anxiety themes for more personalized and effective clinical treatment strategies.

5. Limitations and Future Work

This study has some limitations that should be taken into account when interpreting the results. Firstly, the dataset used in this study may not reveal general patterns apply to all individuals experiencing social anxiety, as social anxiety issues can vary based on regional, cultural, and socio-economic factors. Additionally, the dataset is primarily taken from a single social media source i.e. Reddit, a larger dataset incorporating multiple social media platforms and a broader range of user 11demographics could provide more comprehensive results and reveal more general patterns of social anxiety. Future research may include multiple data sources to improve the robustness and generalizability of the results. By addressing these limitations in future studies, we will try to enhance the understanding of social anxiety across diverse populations in future.

References

1. H. R. Winter, A. R. Norton, J. L. Burley, and B. M. Wootton, "Remote cognitive behaviour therapy for social anxiety disorder: A meta-analysis," *J Anxiety Disord*, vol. 100, p. 102787, 2023, doi: 10.1016/j.janxdis.2023.102787.
2. E. Feldtmann, P. Pointdujour, C. Bellido, and C. Preuss, "Diagnosis and Management of Social Anxiety Disorder," taylorfrancis.comE Feldtmann, P Pointdujour, C Bellido, C PreussAnxiety, Gut Microbiome, and Nutraceuticals, 2023•taylorfrancis.com, pp. 49–68, Jan. 2023, doi: 10.1201/9781003333821-3/DIAGNOSIS-MANAGEMENT-SOCIAL-ANXIETY-DISORDER-ERIK-FELDTMANN-PIER-POINTDUJOUR-CARLOS-BELLIDO-CHARLES-PREUSS.
3. A. Jaiswal, S. Manchanda, … V. G.-J. of family, and undefined 2020, "Burden of internet addiction, social anxiety and social phobia among University students, India," journals.lww.comA Jaiswal, S Manchanda, V Gautam, AD Goel, J Aneja, PR RaghavJournal of family medicine and primary care, 2020•journals.lww.com, Accessed: May 21, 2025. [Online]. Available: https://journals.lww.com/jfmpc/fulltext/2020/09070/Burden_of_internet_addiction,_social_anxiety_and.74.aspx
4. M. Rezaeian, M. Akbari, … A. S.-… of occupational health, and undefined 2020, "Anxiety, social phobia, depression, and suicide among people who stutter; a review study," johe.rums.ac.irM Rezaeian, M Akbari, AH Shirpoor, Z Moghadasi, Z Nikdel, M HejriJournal of occupational health and epidemiology, 2020•johe.rums.ac.ir, Accessed: May 21, 2025. [Online]. Available: https://johe.rums.ac.ir/browse.php?a_id=385&sid=1&slc_lang=en&html=1
5. M. Asher, I. A.-J. of clinical psychology, and undefined 2018, "Gender differences in social anxiety disorder," Wiley Online LibraryM Asher, IM AderkaJournal of clinical psychology, 2018•Wiley Online Library, vol. 74, no. 10, pp. 1730–1741, Oct. 2018, doi: 10.1002/JCLP.22624.
6. E. Stănculescu, M. G.-T. and Informatics, and undefined 2022, "Social media addiction profiles and their antecedents using latent profile analysis: The contribution of social anxiety, gender, and age," ElsevierE Stănculescu, MD GriffithsTelematics and Informatics, 2022•Elsevier, Accessed: May 21, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0736585322001125
7. K. Ranta, M. Inkinen, … E. L.-… of E. and, and undefined 2022, "Adolescents' interpersonal cognition and self-appraisal of their own anxiety in an imagined anxiety-provoking classroom presentation scenario: Gender differences," researchportal.tuni.fiK Ranta, M Inkinen, E Laakkonen, HR Ståhl, N Junttila, PM NiemiEuropean Journal of Education and Psychology, 2022•researchportal.tuni.fi, doi: 10.32457/ejep.v15i2.1969.
8. Z. W. M. Tse, S. Emad, M. K. Hasan, I. V. Papathanasiou, I. Rehman, and K. Y. Lee, "School-based cognitive-behavioural therapy for children and adolescents with social anxiety disorder and social anxiety symptoms: A systematic review," journals.plos.orgZWM Tse, S Emad, MK Hasan, IV Papathanasiou, I Rehman, KY LeePlos one, 2023•journals.plos.org, vol. 18, no. 3 March, Mar. 2023, doi: 10.1371/JOURNAL.PONE.0283329.
9. M. Rizwan and J. Demšar, "Prevalent Frequency of Emotional and Physical Symptoms in Social Anxiety using Zero Shot Classification: An Observational Study," CLPsych 2024 - 9th Workshop on Computational Linguistics and Clinical Psychology, Proceedings of the Workshop, pp. 145–152, 2024.
10. C. A. Stamatis et al., "Prospective associations of text-message-based sentiment with symptoms of depression, generalized anxiety, and social anxiety," Wiley Online LibraryCA Stamatis, J Meyerhoff, T Liu, G Sherman, H Wang, T Liu, B Curtis, LH Ungar, DC MohrDepression and anxiety, 2022•Wiley Online Library, vol. 39, no. 12, pp. 794–804, Dec. 2022, doi: 10.1002/DA.23286.
11. J. M. Whealin, J. J. Saleem, B. Vetter, J. Roth, and J. Herout, "Development and cross-sectional evaluation of a text message protocol to support mental health well-being.," psycnet.apa.orgJM Whealin, JJ Saleem, B Vetter, J Roth, J HeroutPsychological Services, 2023•psycnet.apa.org, 2021, doi: 10.1037/ser0000601.

12. R. da L. Dias et al., "The effectiveness of CBT-based daily supportive text messages in improving female mental health during COVID-19 pandemic: results from the Text4Hope program," frontiersin.orgRL Dias, R Shalaby, B Agyapong, W Vuong, A Gusnowski, S Surood, AJ GreenshawFrontiers in Global Women's Health, 2023•frontiersin.org, vol. 4, p. 1182267, 2023, doi: 10.3389/FGWH.2023.1182267/FULL.
13. L. Ren, H. Lin, B. Xu, S. Zhang, … L. Y.-J. medical, and undefined 2021, "Depression detection on reddit with an emotion-based attention network: algorithm development and validation," medinform.jmir.orgL Ren, H Lin, B Xu, S Zhang, L Yang, S SunJMIR medical informatics, 2021•medinform.jmir.org, Accessed: May 21, 2025. [Online]. Available: https://medinform.jmir.org/2021/7/e28754
14. S. Zanwar, D. Wiechmann, Y. Qiao, E. K.- medRxiv, and undefined 2023, "MANTIS at# SMM4H 2023: Leveraging Hybrid and Ensemble Models for Detection of Social Anxiety Disorder on Reddit," medrxiv.orgS Zanwar, D Wiechmann, Y Qiao, E KerzmedRxiv, 2023•medrxiv.org, doi: 10.1101/2023.12.05.23299439.ABSTRACT.
15. R. Chiong, G. Budhi, S. Dhakal, F. C.-C. in B. and, and undefined 2021, "A textual-based featuring approach for depression detection using machine learning classifiers and social media texts," ElsevierR Chiong, GS Budhi, S Dhakal, F ChiongComputers in Biology and Medicine, 2021•Elsevier, Accessed: May 21, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010482521002936
16. J. Kim, J. Lee, E. Park, J. H.-S. reports, and undefined 2020, "A deep learning model for detecting mental illness from user content on social media," nature.comJ Kim, J Lee, E Park, J HanScientific reports, 2020•nature.com, vol. 10, p. 11846, 2020, doi: 10.1038/s41598-020-68764-y.
17. D. Meshi, M. E.-A. Behaviors, and undefined 2021, "Problematic social media use and social support received in real-life versus on social media: Associations with depression, anxiety and social isolation," ElsevierD Meshi, ME EllithorpeAddictive Behaviors, 2021•Elsevier, 2021, doi: 10.1016/j.addbeh.2021.106949.
18. Malik, S., Iftikhar, A., Tauqeer, F. H., Adil, M., & Ahmed, S. (2022). A Systematic Literature Review on Leukemia Prediction Using Machine Learning. *Journal of Computing & Biomedical Informatics*, *3*(02), 104-123.
19. A. Bedaso, J. Adams, W. Peng, and D. Sibbritt, "The relationship between social support and mental health problems during pregnancy: a systematic review and meta-analysis," SpringerA Bedaso, J Adams, W Peng, D SibbrittReproductive health, 2021•Springer, vol. 18, no. 1, p. 162, Dec. 2021, doi: 10.1186/S12978-021-01209-5.
20. E. O'Day, R. H.-C. in H. B. Reports, and undefined 2021, "Social media use, social anxiety, and loneliness: A systematic review," ElsevierEB O'Day, RG HeimbergComputers in Human Behavior Reports, 2021•Elsevier, Accessed: May 21, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S245195882100018X
21. V. Manova, F. Grosso, B. Khoury, and F. Pagnini, "Social anxiety: topics and emotions shared on Reddit before and during the coronavirus pandemic," SpringerV Manova, F Grosso, B Khoury, F PagniniCurrent Psychology, 2024•Springer, vol. 43, no. 32, pp. 26608–26617, Aug. 2024, doi: 10.1007/S12144-024-05891-Z.
22. M. Lyvers, A. Salviani, S. Costan, and F. A. Thorberg, "Alexithymia, narcissism and social anxiety in relation to social media and internet addiction symptoms," Wiley Online LibraryM Lyvers, A Salviani, S Costan, FA ThorbergInternational Journal of Psychology, 2022•Wiley Online Library, vol. 57, no. 5, pp. 606–612, Oct. 2022, doi: 10.1002/IJOP.12840.
23. U. Lokala et al., "A computational approach to understand mental health from reddit: knowledge-aware multitask learning framework," ojs.aaai.orgU Lokala, A Srivastava, TG Dastidar, T Chakraborty, MS Akhtar, M Panahiazar, A ShethProceedings of the International AAAI Conference on Web and Social Media, 2022•ojs.aaai.org, 2022, Accessed: May 21, 2025. [Online]. Available: https://ojs.aaai.org/index.php/ICWSM/article/view/19322
24. Abbas, F., Iftikhar, A., Riaz, A., Humayon, M., & Khan, M. F. (2024). Use of Big Data in IoT-Enabled Robotics Manufacturing for Process Optimization. Journal of Computing & Biomedical Informatics, 7(01), 239-248.
25. D. Low, L. Rumker, T. Talkar, J. Torous, … G. C.-J. of medical, and undefined 2020, "Natural language processing reveals vulnerable mental health support groups and heightened health anxiety on reddit during covid-19:

Observational study," jmir.orgDM Low, L Rumker, T Talkar, J Torous, G Cecchi, SS GhoshJournal of medical Internet research, 2020• jmir.org, Accessed: May 21, 2025. [Online]. Available: https://www.jmir.org/2020/10/e22635/

26. W. Huang et al., "An empirical study of llama3 quantization: From llms to mllms," SpringerW Huang, X Zheng, X Ma, H Qin, C Lv, H Chen, J Luo, X Qi, X Liu, M MagnoVisual Intelligence, 2024•Springer, Apr. 2024, doi: 10.1007/S44267-024-00070-X.
27. V. Pant, R. Sharma, S. K.-A. in Networks, undefined Intelligence, and undefined 2024, "An overview of stemming and lemmatization techniques," taylorfrancis.comVK Pant, R Sharma, S KunduAdvances in Networks, Intelligence and Computing, 2024• taylorfrancis.com, pp. 308–321, Apr. 2024, doi: 10.1201/9781003430421-31/OVERVIEW-STEMMING-LEMMATIZATION-TECHNIQUES-VINAY-KUMAR-PANT-RUPAK-SHARMA-SHAKTI-KUNDU.
28. M. Franke and J. Degen, "The softmax function: Properties, motivation, and interpretation," 2023, Accessed: May 21, 2025. [Online]. Available: https://osf.io/preprints/psyarxiv/vsw47/
29. Iftikhar, A., Elmagzoub, M. A., Shah, A. M., Al Salem, H. A., ul Hassan, M., Alqahtani, J., & Shaikh, A. (2023). Efficient Energy and Delay Reduction Model for Wireless Sensor Networks. Comput. Syst. Sci. Eng., 46(1), 1153-1168.