‘I just need to vent’ – A Text Mining Study Analysing Reddit Discourse on Obsessive-Compulsive-Disorder on r/OCD using BERT-based models

Leona Weise

Faculty of Behavioural, Management and Social Sciences, University of Twente

Master Thesis

First supervisor: Peter ten Klooster

Second supervisor: Jorge Piano Simões

July 10, 2025

Abstract

Obsessive-compulsive disorder is a highly prevalent and chronic mental health condition, characterized by persistent and distressing obsessions and time-consuming compulsions, causing significant impairment in daily functioning. Yet, research on the lived experience of individuals with OCD remains limited. As people increasingly seek out anonymous online spaces such as Reddit to discuss their experiences with OCD, these firsthand accounts present a novel opportunity for insight. This study leveraged transformer-based BERT models (*BERTopic*, *roBERTa-base-sentiment*, *roBERTa-base-emotion*) to conduct topic modelling, sentiment analysis and emotion recognition on 5,083 posts from the *r/OCD subreddit* that were scraped between February and April 2025 using the *Hot* sorting algorithm. Topic modelling could identify both well-established and lesser researched aspects of OCD across 34 topics. Most topics represented known obsessions and compulsions that could be attributed to well-known OCD dimensions, such as sexuality and contamination. Other topics discussed general OCD concepts (i.e., obsessions), treatment, social impact, and substance use. Moreover, analysis captured understudied aspects of OCD, namely its intersections with pregnancy and social media and its impact on family. The majority of these topics were overwhelmingly negative in sentiment and contained fear and sadness as the most prominent emotions, underscoring the distressing nature of OCD. These results demonstrate that transformer-based text mining can effectively validate established clinical knowledge and capture novel insights, identifying unmet OCD needs from Reddit data. This method offers a scalable, low-preprocessing framework to analyse Reddit data for clinical insight that can be adapted for future research.

A Text Mining Study Analysing Reddit Discourse on Obsessive-Compulsive-Disorder on r/OCD using BERT-based models

Obsessive-compulsive disorder (OCD) is a debilitating psychiatric condition that can significantly impair an individual's functioning and quality of life (Reddy et al., 2017; Ruscio et al., 2010). It is characterized by distressing obsessions and compulsions, which are often highly time-consuming (Stein et al., 2019). Obsessions are unwanted, repetitive intrusive thoughts, imagery or urges (APA, 2022; Reddy et al., 2017). Common dimensions of obsessions include contamination, harm, sexuality, religious ideas, and achieving a *just right* feeling or symmetry (Harrison et al., 2017; Homonoff & Sciutto, 2019; Stein et al., 2019). To neutralize or reduce the anxiety evoked by obsessions, or to prevent an imagined undesirable event from occurring, compulsions are performed (APA, 2022). Compulsions are repetitive behavioural or mental acts, such as checking, washing or counting (Harrison et al., 2017). They are usually excessive, subject to strict rules, and not realistically related to the consequence they aim to prevent (APA, 2022). For example, a fear of contamination may cause the urge to follow excessive hygiene rituals, such as extensive handwashing to the point of skin lesions. However, OCD is a heterogeneous disorder and may present very differently across clinical cases (Ponzini & Steinman, 2022). It affects 1-3% of the general population in their lifetime, making it the fourth most common psychiatric illness (Gomes et al., 2023). Despite being highly prevalent, most sufferers only receive appropriate treatment years after onset of symptoms and knowledge about the lived experience of OCD is limited (Fennell & Liberato, 2007; Reddy et al., 2017; Steinberg et al., 2017; Ziegler et al., 2021).

OCD is subject to stigmatization across interpersonal, clinical and cultural domains (Ponzini & Steinman, 2022; Wheaton et al., 2016). Fearing dismissal, rejection, or being perceived as weak or dangerous, individuals with OCD often conceal their symptoms, leading to social isolation and treatment delay that can worsen symptoms and result in a chronic state

(Bathje & Pryor, 2011; Cipolla et al., 2024; Goodman et al., 2014; Schofield & Ponzini, 2020). OCD remains misunderstood even among clinicians, with misidentification rates as high as 53% (Perez et al., 2022), in part due to bias against certain dimensions of obsessions, such as harm or sexuality (Glazier et al., 2015; Hirschtritt et al., 2017; Homonoff & Sciutto, 2019). It has been established that media representations often reinforce these misconceptions through stereotypical portrayals or trivialization on platforms like TikTok, perpetuating public misunderstanding of the disorder (Fennell & Boyd, 2014; Woods et al., 2023).

Yet, in the same digital landscape, individuals with OCD seek out others with the disorder in dedicated online forums that offer anonymity and peer connection, finding relief in connecting over shared experiences. (Fennell & Liberato, 2007). With over 5 billion users worldwide, social media (SM) platforms can offer accessible, low-threshold environments for peer interaction that mirror the supportive dynamics found in traditional self-help groups (Kemp, 2024; Sidani et al., 2016). SM has the distinctive ability to foster strong connections between individuals regardless of location, age, and gender, forming online 'communities' bound by shared experiences (Giordani & Silva, 2021). Members can provide and receive support and information about treatment options, symptoms, and other personal experiences (McCormack, 2010; Tan et al., 2021).

Among leading social media platforms, Reddit stands out as a particularly significant space for mental health discourse due to its unique forum structure. Reddit has evolved into one of the biggest social media platforms, with approximately 108.1 million daily active users and over 138,000 active topic-specific communities referred to as *subreddits* (Proferes et al., 2021; Reddit, Inc., 2025). Subreddits are created and moderated by users themselves, complete with their own rules and language features (Medvedev et al., 2018; Proferes et al., 2021). They offer a secluded online space, and leave much control to thread moderators, some of whom are licensed mental health professionals, to set and enforce rules.

Reddit fosters conversations about stigmatized mental illness by prioritizing user anonymity through pseudonymous accounts and platform norms that dissuade users from disclosing personal information (Proferes et al., 2021). Given the stigmatization associated with OCD, the anonymity afforded by Reddit's platform architecture is instrumental in facilitating candid discussion, including disclosures of distressing or taboo symptoms (De Choudhury & De, 2014; Sit et al., 2022). Such self-disclosure and disinhibition are found to enhance emotional expression and increase the likelihood of receiving supportive feedback (De Choudhury & De, 2014; Fennell & Liberato, 2007).

The platform generally sets itself apart from others (e.g., Instagram or Facebook) through a focus on topic rather than user identity (Brown et al., 2018; Sit et al., 2022; Zhang et al., 2017). This structure of topic-specific communities makes it uniquely suited for academic research, as relevant data can efficiently be located and extracted (Proferes et al., 2021). In comparison to other well-known SM platforms, Reddit posts tend to be more textual and discussion-focused in nature without strict character limits. This results in discourse that is both qualitatively rich and high in quantity, lending itself well to academic study (Proferes et al., 2021). In addition, Reddit fosters data collection for research through a publicly available Application Programming Interface (API).

There are several subreddits specifically dedicated to topics associated with OCD. The largest is *r/OCD*, a subreddit "dedicated to discussion, articles and support regarding OCD" (r/OCD, n.d.; r/OCD, n.d., para. 1) where users have created a space to *vent* about their experiences and discuss symptoms and treatment options. It was created on 30 March 2009, and has since grown to over 270,000 members (r/OCD, June 27, 2025). Traditionally, discourse among those affected by OCD has occurred in intimate settings closed off to the public. On Reddit, r/OCD provides unprecedented access to real-life, unsolicited discussions about the lived experiences of OCD.

Despite increasing availability of online discussions of OCD, no research to date has examined how individuals with OCD discuss the disorder among themselves. These conversations bear potential to provide a comprehensive understanding of the heterogeneity of OCD and its impact on individual lives. Van Schalkwyk et al. (2017) highlighted the fundamental need for approaches that incorporate patient narratives to allow for open-ended exploration of the lived experience of OCD. Such insight may complement and expand existing knowledge about comorbidities, OCD subtypes as well as shape treatment protocols and existing scales, which have largely been informed by the expertise of clinicians, rather than patient perspectives (Van Schalkwyk et al., 2016; Subramaniam et al., 2014). Online discussions among people with OCD can present an opportunity to access more diverse and possibly highly stigmatized or 'taboo' perspectives by those who are most affected by the disorder. The bulk of existing research on individual experiences with OCD to date has used traditional quantitative and qualitative methodologies, such as questionnaires (e.g., Patel et al., 2017), interviews (e.g., Kohler et al., 2018; Keyes et al., 2017; van Schalkwyk et al., 2016; Sravanti et al., 2022) and self-reported observation (Zisler et al., 2024). Each of these methods is subject to limitations such as selection bias, hindsight bias, recall bias or obtrusiveness, which may be especially detrimental considering the shame and stigmatization surrounding OCD symptoms. In contrast, observing real-life interactions on r/OCD, an anonymous, user-driven space, can reduce different types of bias and provide access to unsolicited patient accounts.

Several studies have successfully used posts on social media to gain more insight into mental health disorders and better understand the experiences of people with different mental health issues such as OCD (Al-Haider et al., 2024; Boettcher, 2021; De Choudhury & De, 2014; 2023; Park et al., 2018; Sit et al., 2022; Kim et al.; Woods et al., 2023). As subreddits contain vast amounts of unstructured data in the form of digitized text, however, the quantity

of text material available quickly exceeds the scope and feasibility of traditional methods of content analysis, such as manual coding or grounded theory qualitative analysis, as they are labor intensive (Antons et al., 2020). Text mining has emerged as a promising method to overcome some capacity limitations of manual content analysis. According to Antons et al. (2020, p.330), "the case for text mining becomes stronger the larger the text corpus and the less accessible it is to manual content analytical techniques".

Text mining refers to the extraction of unknown, implicit patterns and information from a large corpus of unstructured textual data using computer algorithms (Hassani et al., 2020). Through text mining, a higher volume of data can be analysed, providing a more representative analysis of OCD discussions on SM than would be feasible through manual analysis. Three well-established text mining applications are topic modelling, sentiment analysis and emotion recognition. Topic modelling refers to the autonomous extraction of topics from text (Albalawi et al., 2020). Sentiment analysis estimates the polarity of attitudes, views, feelings and opinions expressed, while emotion recognition means the identification of specific human emotions, such as sadness, fear, or joy in a text (Mohammad et al., 2018; Nandwani & Verma, 2021; Plutchik, 1980; Yadav & Vishwakarma, 2020). Recent advances in natural language processing have brought forth transformer-based models, so-called bidirectional encoder representations from transformers (BERT) models, for topic modelling, sentiment analysis, and emotion recognition that possess improved ability to process contextual information and nuances in language and pick up on sarcasm (Egger & Yu, 2022).

To date, no studies have explored the Reddit discourse on OCD for its content, such as topics, sentiments and emotions using text mining methods and there remains a lack of research into the actual discourse within such communities. Given the documented stigmatization and widespread lack of understanding of OCD, taking advantage of this novel access to discourse on r/OCD could provide unique insights into the lived experience of

people with OCD. Therefore, this study used advanced transformer-based topic modelling, sentiment analysis and emotion recognition to examine the discourse in the r/OCD subreddit. This was done by addressing the following research questions:

RQ1: What topics are discussed on r/OCD?

RQ2: What sentiment and which emotions are expressed in discussions on r/OCD?

Method

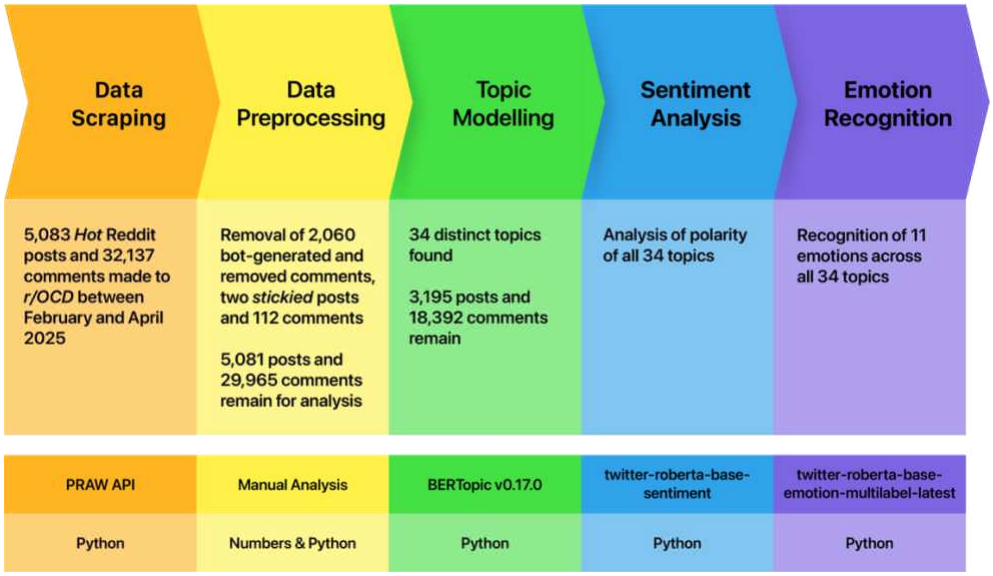
Study Design

Using an exploratory data-driven approach, the present study leveraged unsupervised transformer-based text mining techniques to analyse posts and comments made to the r/OCD subreddit. Text mining studies using SM data typically consist of several consecutive steps, generally referred to as a text mining pipeline, starting with the scraping of postings up until the specific text-mining methods and analyses. The text-mining pipeline for the current study is visualized in Figure 1. The Python script for the subsequent procedure can be found on GitHub[1]. This study received ethical approval from the University of Twente Humanities & Social Sciences Ethics Committee.

[1] https://github.com/LeoWeise/Master-Thesis-OCD

Figure 1

Text Mining Pipeline



Software

This study used Python version 3.10 in combination with the PyCharm 2024.3.2 Community Edition IDE for script development and execution.

Data scraping

This study focused on original posts, including title, post body and all comments, made to r/OCD. A summary of the moderator-curated information and rules for r/OCD can be found on GitHub[2]. First, a dataset was accumulated by scraping Reddit posts and comments. Data scraping refers to the process by which data is extracted from an online platform and stored in a structured format for further analysis (Glez-Peña et al., 2014). The *Python Reddit API Wrapper 7.7.1 (PRAW)*[3] library was used to access Reddit's API. An API is designed to facilitate data access in line with technical boundaries and ethical data collection standards (Proferes et al., 2021). PRAW contains custom web-scraping scripts and offers five post-

[2] https://github.com/LeoWeise/Master-Thesis-OCD
[3] https://praw.readthedocs.io/en/stable/

sorting algorithms, namely *Best*, *Hot*, *New*, *Top* and *Rising* (Lundblade, 2023). The present study used the Hot algorithm to scrape posts prioritized by recent community engagement (Lundblade, 2023). It elevates posts which are recent and popular, indicated by vote score (upvotes minus downvotes) and a logarithmic function, according to which votes are weighted higher the closer they occur to the time of submission (Lundblade, 2023; Salihefendic, 2015). For example, the first 10 upvotes are assigned the same weight as the next 100, which in turn are weighed the same as the next 1000 (Salihefendic, 2015). This approach reflects active and organically evolving discourse, aligning with the research focus on prominent themes.

Reddit's API limits extraction to 1000 posts along with all associated comments. Thus, batches of 1000 *Hot* posts including all comments were scraped between 06.02.2025 and 11.04.2025 resulting in 5,083 unique posts and 32,137 comments. The resulting data was merged and stored in a JSON file.

Data Preprocessing

Noise was identified through manual analysis after the first few topic model runs. In the context of this study, noise refers to posts and comments that are not organic user contributions and may distort analysis of genuine user discourse. Accordingly, posts made by moderators, which are *pinned* to the top of the r/OCD *feed*, bot-generated and *removed* comments were filtered out. In total, two stickied posts and 112 associated comments, 1,174 *Spoiler/NSFW* bot comments, 541 *Suicide Prevention* bot comments, and 345 removed comments were deleted, and 5,081 posts and 29,965 comments remained. Minimal preprocessing was performed, as BERT has demonstrated great performance on raw textual data (Rezapour, 2024). Research has even found the accuracy of BERT to be reduced by preprocessing on some NLP tasks, as important context may accidentally be discarded (Rezapour, 2024). By retaining nearly all data, the chance for human error was reduced and

the workflow simplified (Egger & Yu, 2022). To preserve the full context of user interactions, each Reddit post, including title, body and associated comments were concatenated into one single document, resulting in one input sequence per post.

Data Mining and Analysis

Text mining techniques convert unstructured data from textual into numerical form, allowing the application of mathematical algorithms (Antons et al., 2020). Three text mining techniques, namely topic modelling, sentiment analysis and emotion recognition were applied to the data using BERT-based models. BERT is a simple and empirically powerful language representation model trained on broad data (Bommasani et al., 2021; Devlin et. al, 2019; Liu et al., 2019; Rezapour, 2024). It generates deep bidirectional word encodings by processing text with both the left and right context across all layers of its architecture, enabling it to capture meaning and context more accurately (Devlin et. al., 2019, Grootendorst, 2022). BERT has demonstrated strong semantic understanding by generating contextual word- and sentence vector representations that cluster similar text closely in vector space (Grootendorst, 2022). This sophisticated processing may enable BERT to outperform manual coding, both in terms of efficiency, and in accuracy and stability, by reducing potential for error (Egger & Yu, 2022; Sukumar et al., 2024; Zong et al., 2021). Its performance on several measures evaluating language understanding reflects this strength, achieving high scores on diverse performance metrics (Devlin et. al, 2019).

BERT can understand, extract and organize information even from complex written text outperforming both traditional NLP approaches and newer GPT models in some language understanding and classification tasks (Chowdhary, 2020; Jurafsky & Martin, 2025; Wang et al., 2024; Zong et al., 2021). This capability is particularly useful for Reddit posts, which present unique linguistic challenges, including informal language, community-specific jargon, abbreviations, and emotionally charged content. In addition, text volume per post can

be substantial due to the platform's liberal character limitations, providing greater complexity, nuance and rich context.

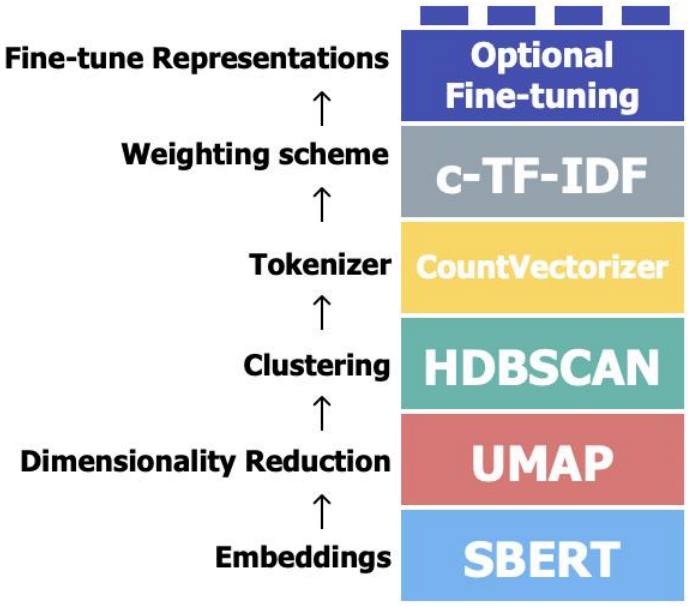
Topic Modelling

Latent themes in r/OCD were identified through topic modelling. The resulting topic representations were used to derive and explore underlying themes from posts in r/OCD.

Topic Modelling with BERTopic. BERTopic has emerged as a powerful and accurate transformer-based method. Outperforming three other prominent topic models in comparison, its algorithm demonstrates strong performance in most variations of topic modelling, whereas others are often limited to one (Egger & Yu, 2022; Grootendorst, 2022). In addition, BERTopic demonstrates stability in terms of two widely used metrics, namely topic coherence, meaning that the topics reasonably emulate human judgement, and topic diversity, referring to the percentage of unique words per topic (Grootendorst et al., 2022). It was chosen for its suitability for unstructured textual social media data such as Reddit posts (Egger & Yu, 2022). The BERTopic pipeline consists of six modules, each of which can be adapted to the data at hand, a structure referred to as *Build Your Own Topic Model* (See Figure 2) (Grootendorst, 2022).

Figure 2

General BERTopic Pipeline



Build your own topic model. In this study, the default configuration of BERTopic (Figure 2) was optimized for the dataset (English) using the library *bertopic* by Grootendorst (2022). First, *Sentence-BERT (SBERT),* which efficiently creates robust sentence level embeddings, was used (Reimers & Gurevych, 2019). Specifically *all-MiniLM-L12-v2*[4] was implemented, which is five times faster than the base model while retaining good quality. Next, *Uniform Manifold Approximation and Projection (UMAP)* was used to reduce the dimensionality of the embeddings, allowing *HBDSCAN* to identify clusters (Grootendorst, 2022). UMAP has shown to improve HBDSCAN clustering in terms of accuracy and efficiency (Allaoui et al., 2020; Grootendorst, 2022). Since UMAP is a stochastic algorithm, it was adjusted to be deterministic for reproducibility by assigning a random seed number. HBDSCAN clusters use a soft-clustering approach, allowing for noise to be modelled as

[4] https://huggingface.co/microsoft/MiniLM-L12-H384-uncased

outliers instead of assigning it to clusters, leading to better topic representations for the remainder of the corpus (Grootendorst, 2020). Both UMAP and HBDSCAN parameters were adjusted to allow for smaller clusters (n_neighbors=10, n_components=3; min_cluster_size=15, min_samples=10). Next, *CountVectorizer* was used to generate topics from the embeddings clustered in the prior step. Additionally, stop words were removed (English) and an n-gram range was added (1, 2). Lastly, class-based *term frequency-inverse document frequency (c-TF-IDF)* was used to create topic representations. TF-IDF assumes that the more frequent a term is, the less information it provides to a document. The c-TF-IDF approach modifies traditional TF-IDF to work at the cluster level rather than document level, allowing for better representation of topic-word distributions by measuring how important words are to entire topic clusters (Grootendorst, 2022). In terms of fine-tuning, multiple steps were taken. First, *KeyBERTInspired* was used to create meaningful keyword topic representations. Next, topics were internally reduced to a number of 35 topics, consolidating similar topics, but allowing for a large variety of topics to remain. This allowed for small clusters, representing niche, but distinct topics, to remain in the final topic selection. Multiple means of visualization were selected, namely an intertopic distance map, bar charts, hierarchical clustering and a heatmap representation.

Topic Coherence. In this study, C_v and NPMI coherence measures were used in combination with the id2word dictionary. The C_v coherence measure ranges from 0-1 and combines different similarity measures and segmentation strategies to achieve strong correlation with human judgement (Röder et al., 2015). NPMI coherence measure is based on word co-occurrence and has been found to correlate well with human judgement, ranging from -1 to 1 (Lau et al., 2014; Röder et al., 2015). Higher scores indicate better coherence.

Topic labelling. Topic labels were created based on all documents and keywords. First, via Python using OpenAI's GPT-4o by iteratively tailoring a prompt until the labels

were deemed sufficiently congruent with the topics. In addition, the topics were human-labelled based on the AI-generated labels and detailed manual inspection. The prompt is available on GitHub[5].

Sentiment Analysis and Emotion Recognition

The generated topics were analysed for underlying attitudes and affective tone by conducting sentiment analysis (SA) and emotion recognition (ER) using two unsupervised transformer-based models derived from the Robustly optimized BERT (RoBERTa) approach. RoBERTa improves upon the original BERT model by training on longer sequences and larger batches of data, achieving state-of-the-art results on several benchmark evaluations (Liu et al., 2019). It was further pretrained on 58 million tweets by *Cardiff NLP,* resulting in Twitter-RoBERTa-base, which was fine-tuned for sentiment analysis and emotion recognition using the *TweetEval* benchmark (Barbieri et al., 2020). *TweetEval* is a new assessment framework for Twitter-specific classification tasks, such as sentiment analysis, emotion recognition, offensive language detection, emoji prediction and irony detection (Barbieri et al., 2020). While Reddit data tends to contain longer, more complex sequences than Twitter data, these models were chosen for accessibility and their ability to handle linguistic nuances specific to SM discourse (Barbieri et al., 2020; Liu et al., 2019).

Twitter-roBERTa-base for Sentiment Analysis. The latest version of *twitter-roberta-base-sentiment*[6] was used to examine the polarity of each document (post including title, body and comments) (Barbieri et al., 2020). It categorizes text as positive, neutral, or negative (Barbieri et al., 2020; Camacho-Collados et al., 2022).

Twitter-roBERTa-base for Emotion Recognition. Second, *twitter-roberta-base-emotion-multilabel-latest*[7] was used (Camacho-Collados et al., 2022). It conducts emotion

[5] https://github.com/LeoWeise/Master-Thesis-OCD
[6] https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment
[7] https://huggingface.co/cardiffnlp/twitter-roberta-base-emotion-multilabel-latest

recognition by classifying posts according to 11 emotional categories, namely anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, surprise and trust (Camacho-Collados et al., 2022). The categories were derived from the *SemEval-2018 Task 1*, which was based on Plutchik's wheel of emotions with three additional emotions, namely love, optimism and pessimism added based on observations (Mohammad et al., 2018; Plutchik, 1980). Emotion recognition was used to complement the sentiment analysis by providing a more fine-grained view. Analysis was conducted on the 50 most representative documents from each topic, or all, if the number of total documents was below 50. For each post, the models conducted the analysis of sentiment and probability scores for all 11 emotions simultaneously, then the mean probability sentiment and emotion scores were calculated for each topic. Additionally, all emotion scores were visualized using a line chart for accessibility.

Results

This study explored the topics, sentiments and emotions expressed in posts on r/OCD. The following presents the findings obtained through topic modelling, sentiment analysis and emotion recognition.

Topic Modelling

BERTopic identified 34 meaningful topics, clustering 3,195 of 5,081 posts (62.9%) into groups of 15 to 683 posts per identified topic. The final 3,195 posts contained 18,392 comments. The remaining 1,886 posts (37.1%) were labelled outliers by the HDBSCAN clustering algorithm, as they were too semantically diverse to be clustered into coherent groups. These posts were excluded to ensure clear, coherent topic representations.

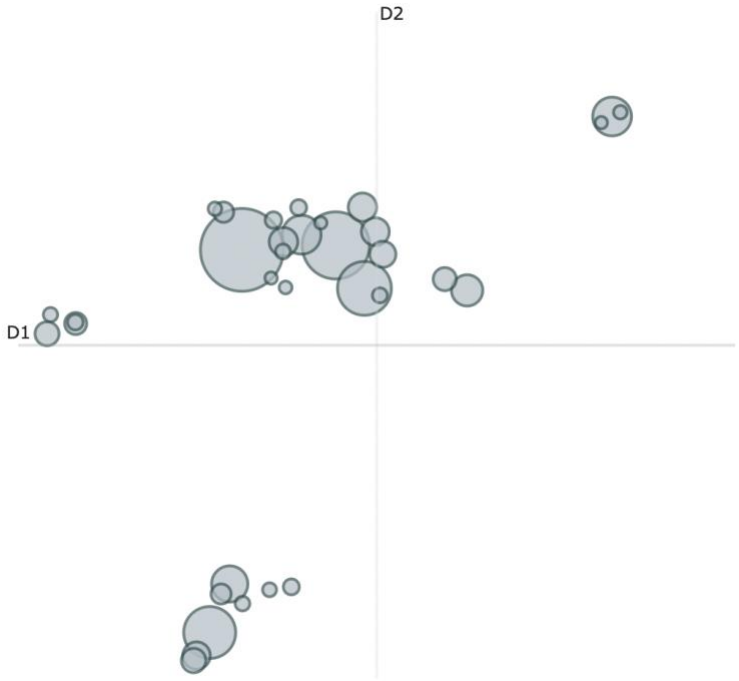
Intertopic Distance

Figure 3 shows a two-dimensional intertopic distance map representing the semantic similarity between the extracted topics. Each circle represents a topic, with the size of the

circle representing the number of posts associated with the topic and their distance indicating how similar they are in content. Thus, clusters of circles represent themes which underlie multiple topics. In total, the final topic representation shows 34 topics, clustered into one large central group and four smaller groups, three of which are peripheral and semantically distinct from the largest cluster. The central cluster contains 17 topics and displays family- and substance related topics clustered around general OCD, contamination-related topics clustered near OCD Medication information, and relationship OCD in close proximity to reassurance-seeking. Adjacent to the central cluster is a small treatment-related two-topic cluster. On the top right of the map, three topics revolving around fear of death, namely existential, health and flight OCD form a distinct group. A cluster on the left comprises four perfectionism-related topics including music, reading and writing, memory hoarding and gaming-related OCD. At the bottom, a cluster is comprised of one three-topic section related to intrusive thoughts, guilt and false memory OCD, while the other half is more loosely spaced and consists of five topics related to compulsions, such as skin picking, dental OCD and numerical OCD. The spatial separation between the topic groups indicates distinct underlying themes and high variability of topics.

Figure 3

Intertopic Distance Map



Note. An interactive version of this figure with topic assignments is available on GitHub. [8]

Topic Coherence

Topic Coherence was calculated after topic reduction. C_v coherence was strong (0.635) and NMPI coherence was moderate (0.116) (Röder et al., 2015).

Topic Overview

Table 1 shows the 34 final topics, including LLM-generated and human- assigned topic label, document count per topic, percentage of the total number of posts, probability, as well as 10 representative keywords. A detailed overview of all Reddit posts and comments by topic, a hierarchical chart, bar chart representations and a heatmap representation can be accessed on GitHub[9].

[8] https://leoweise.github.io/Master-Thesis-OCD/Intertopic Distance Map.html
[9] https://github.com/LeoWeise/Master-Thesis-OCD

Table 1

Labelled Overview of all 34 Topics

Topic	Count	% of Total	Probability	LLM-generated label	human-assigned label	Topic Words
0	683	21.38	0.929	ADHD and OCD Overlap	General OCD characteristics	ocd, disorder, compulsions, intrusive thoughts, therapist, adhd, therapy, obsessions, diagnosed, anxiety
1	449	14.05	0.966	OCD Medication Experiences	OCD medication	zoloft, antidepressants, ocd, prozac, sertraline, seroquel, ssris, ssri, prescribed, medications
2	288	9.01	0.896	Relationship OCD Reassurance Seeking	Relationship and sexuality OCD	ocd, therapy, advice, boyfriend, relationship, reassurance, bf, dating, husband, relationships
3	268	8.39	0.687	Anxiety-Inducing Intrusive Thoughts	Intrusive thoughts	intrusive thoughts, intrusive thought, thoughts, anxiety, therapy, intrusive, anxious, thinking, think, mind
4	150	4.69	0.779	Contamination Fear and Handwashing Rituals	Contamination OCD	washing hands, wash hands, contamination ocd, hand washing, washing, wash, showering, contamination, soap, cleaning
5	149	4.69	0.854	Fear of Death and Existence	Existential OCD	existential ocd, ocd, anxiety, therapy, existential, dying, living, manifestation, death, die
6	130	4.07	0.753	Compulsive Behavior Struggles	Compulsions	ocd compulsions, stop compulsions, doing compulsions, doing compulsion, compulsions just, compulsions, compulsion, compulsion right, compulsion just, lot compulsions
7	97	3.04	0.794	Finding Effective OCD Therapists	OCD therapy	ocd therapist, therapist, good therapist, new therapist, therapy, ocd specialist, therapists, psychiatric consultant, talk therapy, ocd
8	81	2.54	0.666	Sleep-Related OCD Anxiety	Effects of OCD on sleep	sleep ocd, sleep deprived, sleep anxiety, ocd dreams, dreams ocd, getting sleep, falling asleep, insomnia, trying sleep, fall asleep
9	80	2.5	0.690	Religious Scrupulosity and Doubt	Scrupulosity and religious OCD	religious ocd, moral ocd, ocd, religious, religion, spirituality, religions, atheist, believe god, agnostic

Topic	Count	% of Total	Probability	LLM-generated label	human-assigned label	Topic Words
10	77	2.41	0.805	Pet-Related Rabies Fear	Pet-related OCD and Fear of zoophilia	ocd cat, ocd, rabies ocd, pets, pest control, cats, pet, pest, tick bites, cat
11	75	2.35	0.883	False Memory Intrusive Doubts	Memory issues	memory ocd, ocd, false memories, memory syndrome, remember memory, false memory, memories real, memory real, memory false, memories
12	68	2.13	0.668	Food Contamination Fear	Food-related OCD	ocd food, disordered eating, eating disorder, eating disorders, issues eating, ocd, contamination ocd, food poisoning, fear food, eating food
13	59	1.85	0.793	Guilt and Confession Obsessions	Guilt sensitivity and real event OCD	guilt ocd, feel guilt, ocd guilt, feel guilty, hold guilt, extreme guilt, guilt, forgive, guilt shame, confess
14	57	1.78	0.490	Social Media Deletion Anxiety	Social media OCD	memory hoarding, delete things, delete, deleting, hoarding, deleted social, tumblr, instagram, delete app, social media
15	53	1.66	0.928	Starting ERP Therapy	OCD treatment	ocd, exposure therapy, therapy, therapist, doing erp, psych ward, psych wards, started erp, psych, anxiety
16	48	1.5	0.941	Gaming-Related OCD Obsessions	OCD and gaming	ocd video, video games, games, gaming, video game, hobbies interests, hobbies, rpgs, play game, game
17	41	1.28	0.967	Family Contamination Anxiety	Family difficulties with OCD	ocd mom, ocd, contamination ocd, abused parents, understand sister, older sister, sister, mom, abused, parents
18	40	1.25	0.951	Attachment to Inanimate Objects	Fear of damaging possessions	inanimate objects, toys, stuffed animals, objects, stuffed animal, felt, object, feel, couch, toy
19	27	0.85	0.888	Contamination Fear Patterns	OCD themes	ocd themes, ocd, contamination ocd, common themes, themes, theme contamination, certain themes, know themes, themes just, hints themes

Topic	Count	% of Total	Probability	LLM-generated label	human-assigned label	Topic Words
20	25	0.78	0.800	Compulsive Skin Picking	Skin picking	picking scalp, scalp picking, skin picking, picking skin, pick scalp, skin pick, stop picking, picking fingers, scalp, constantly picking
21	25	0.78	0.932	OCD Fear of Pregnancy	Pregnancy-related OCD	ocd pregnancy, pregnancy ocd, pregnant ocd, phobia pregnancy, paranoid pregnant, fear pregnant, fear pregnancy, terrified pregnant, scared pregnant, postpartum anxiety
22	24	0.75	1.0	Obsessive Writing Perfectionism	Reading and writing perfectionism	ocd writing, writing ocd, obsessive writing, reading, struggle read, read properly, struggling rereading, rereading, perfectionism writing, writing novels
23	22	0.69	0.949	POCD Support and Seeking Help	Pedophilic OCD	people pocd, struggled pocd, pocd, pedophile pocd, pocd haven, seek psych, make friends, counselor, friends, willing help
24	22	0.69	0.901	Reassurance-Seeking Compulsion	Reassurance-seeking	getting reassurance, seek reassurance, providing reassurance, ask reassurance, anymore reassurance, reassurance reassurance, reassurance, réassurance, reassurance ai, reassurance isn
25	21	0.66	0.830	Anxiety Around Specific Numbers	Numerical OCD	number ocd, numerical ocd, counting ocd, numbers time, fear number, numbers, numbers numbers, bad numbers, numbers number, odd numbers
26	20	0.63	0.970	Music-Related Obsessive Thoughts	Music OCD and therapeutic music	music therapy, enjoy music, music, symptom obsessive, songs, listen music, thoughts compulsive, listening music, music don, song
27	19	0.59	0.986	Jaw Clenching Anxiety	Dental OCD	teeth ocd, clenching jaw, ocd teeth, jaw clenching, clench jaw, jaw constantly, teeth constantly, clenching constantly, jaw sore, talk dentist
28	18	0.56	0.990	Flight-Related OCD Fears	Flight anxiety	flight anxiety, flight panic, fear flying, plane crash, missed flight, panic attack, plane going, flights, emotional reactions, flight

Topic	Count	% of Total	Probability	LLM-generated label	human-assigned label	Topic Words
29	18	0.56	0.987	Parenting and OCD Concerns	Parenting advice for OCD child	ocd, insists ocd, intelligent ocd, member ocd, therapy, therapist, daughter sudden, daughter, parenting, old daughter
30	16	0.5	0.986	Alcohol-Related OCD Concerns	Alcohol and OCD	alcohol ocd, ocd alcohol, ocd drinking, drinking ocd, worsen ocd, fear alcoholic, alcohol compulsion, alcohol withdrawal, stopping alcohol, quit drinking
31	15	0.47	1.0	Cannabis-Induced OCD Concerns	Cannabis and OCD	root ocd, smoke weed, ridiculous ocd, smoking cannabis, cannabis, smoking weed, ocd clings, weed induced, smoked weed, using weed
32	15	0.47	1.0	Cancer-related Health Anxiety	Health anxiety	health anxiety, cancer, cancer specifically, anxiety, cancer years, like cancer, lung cancer, radiation risks, worried, melanoma
33	15	0.47	1.0	Breathing Fixation OCD	Somatic OCD	breathing ocd, ocd breathing, somatic ocd, breathing obsession, blinking breathing, scared breathing, focusing breathing, exhale exhausting, anxiety time, forgets breath

Note. Each document includes post title, body and comments.

The 34 topics provide a rich, yet difficult to interpret representation of multiple aspects of OCD. To extract insight and structure from such diverse representations, the topics were manually sorted into groups. Most topics (4, 5, 9, 10, 11, 12, 13, 16, 18, 20, 21, 22, 23, 25, 26, 27, 28, 32, 33) represent specific obsessions and compulsions that are attributable to well-researched dimensions of OCD. They are comprised of posts detailing the content of the obsession, occurring feelings and fears, as well as correlated compulsions. For example, some topics can be linked to the contamination dimension, such as fear of food contamination or a cleanliness obsession, while others seem to be about different aspects of the sexuality dimension, such as a fear of being pedophilic or zoophilic. Moreover, other topics represent

manifestations of the perfectionism dimension, pertaining to writing and reading or gaming.

Furthermore, a subset of topics concerns discussions on the treatment of OCD, including information, experiences, or fears related to medication, therapy or specific treatment options such as ERP (1,7, 15). Similarly, another group of topics discusses specific symptoms and aspects of OCD, such as intrusive thoughts, obsession themes, compulsions and reassurance-seeking (3, 6, 19, 24). Two topics address the relationships between substance use and OCD (30, 31). Another group of topics concerns the social effects of OCD in romantic relationships and family contexts (2, 17, 29). Topic 8 centers around discussions about the effect of OCD on sleep. Topic 21 is comprised of posts very specifically discussing the relationship between pregnancy and OCD, while Topic 14 is focused on OCD obsessions and compulsions about social media.

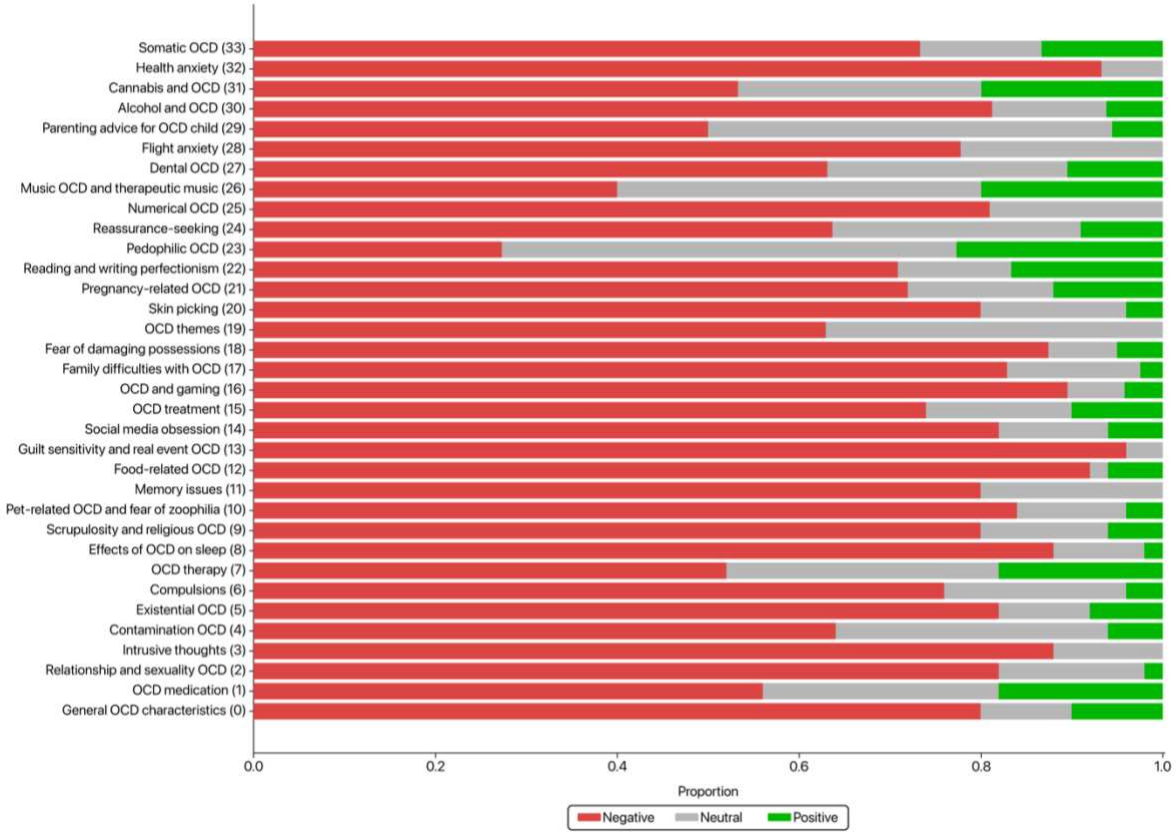
Sentiment Analysis and Emotion Recognition

The majority of documents exhibited a high proportion of negative sentiment, varying between 27% to 98% of posts across all topics with 31 out of 34 topics being predominantly negative in sentiment. Neutral sentiment typically ranged from 2% to 50%, while positive sentiment remained consistently low across all topics, ranging from 2% to 23% of posts. A few topics showed relatively higher levels of positive sentiment. Topic 23 (*Pedophilic OCD*) surprisingly had the highest observed positive sentiment at 23%, followed by Topic 26 (*Music OCD and Therapeutic Music*) and Topic 31 (*Cannabis and OCD*), each at 20%. Topics 3 (*Intrusive Thoughts*), 11 (*Memory Issues*), 13 (*Guilt Sensitivity and Real Event OCD*), 19 (*OCD Themes*), 25 (*Numerical OCD*), 28 (*Flight Anxiety*), and 32 (*Health Anxiety*) contained no posts that were classified as positive. Figure 5 shows the sentiment distribution by topic. All sentiment and emotion scores can be found on GitHub[10].

[10] https://github.com/LeoWeise/Master-Thesis-OCD

Figure 4

Sentiment Polarity Distribution by Topic

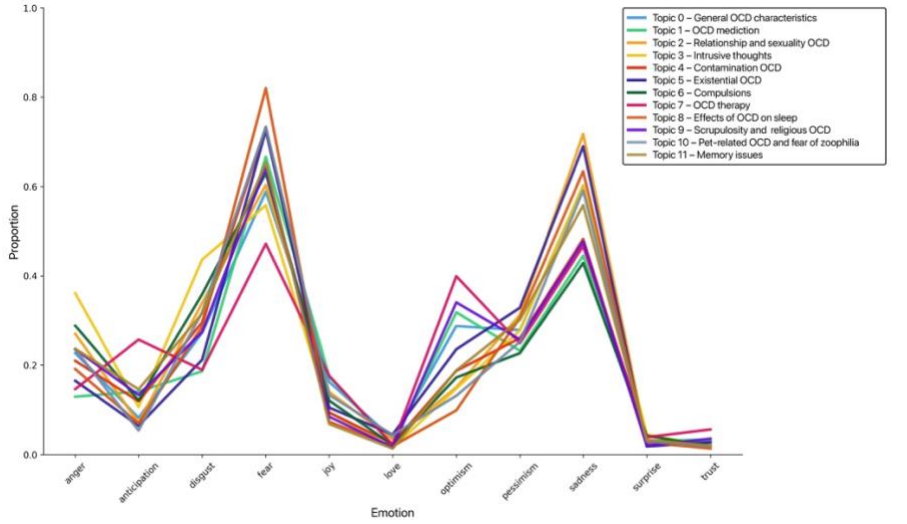


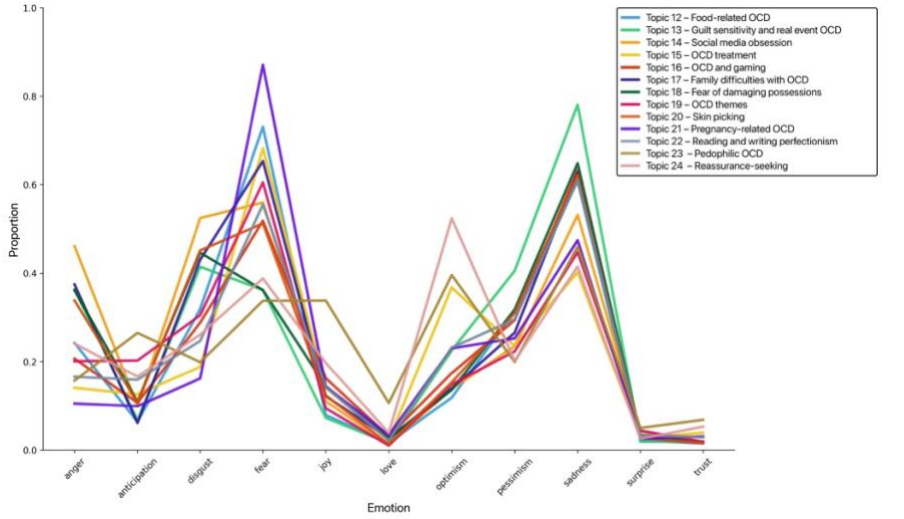
Emotion Recognition

ER revealed that fear (61.8%) and sadness (54.9%) appeared at high intensity across all topics. Out of 34 topics, the dominant emotion was fear in 22 and sadness in 11 topics. Disgust (29.1%) and pessimism (26.5%) were also prevalent across topics. Moderate levels were observed for anger (22.7%) and optimism (22.4%), which was dominant in one topic. Lower intensity emotions included joy (13.9%), anticipation (11.6%), surprise (3.0%), love (2.9%), and trust (2.7%). The probability scores of all 11 emotions across topics are displayed below in Figure 5. While most topics were relatively similar in emotion profiles, topics 23 and 26 stood out with higher joy scores and topic 24 with higher optimism scores. Figure 5 shows all scores per topic and Table 2 displays the most and second most dominant emotion.

Figure 5

Emotion Probabilities across all Topics





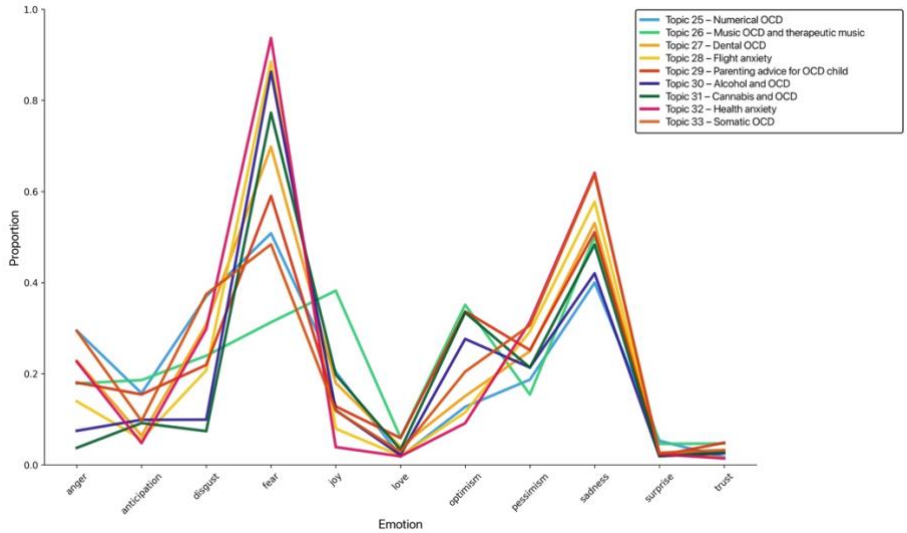


Table 2

Most Dominant and Second Most Dominant Emotions per Topic

Topic		Most Dominant Emotion	Score	Second Most Dominant Emotion	Score
0	General OCD characteristics	sadness	0.602	fear	0.588
1	OCD medication	fear	0.666	sadness	0.445
2	Relationship and sexuality OCD	sadness	0.717	fear	0.604
3	Intrusive thoughts	sadness	0.603	fear	0.558
4	Contamination OCD	fear	0.733	sadness	0.482
5	Existential OCD	fear	0.726	sadness	0.690
6	Compulsions	fear	0.630	sadness	0.429
7	OCD therapy	fear	0.472	sadness	0.466
8	Effects of OCD on sleep	fear	0.820	sadness	0.634
9	Scrupulosity and religious OCD	fear	0.642	sadness	0.477
10	Pet-related OCD and Fear of zoophilia	fear	0.734	sadness	0.591
11	Memory issues	fear	0.654	sadness	0.558
12	Food-related OCD	fear	0.731	sadness	0.608
13	Guilt sensitivity and real event OCD	sadness	0.781	disgust	0.414
14	Social media OCD	fear	0.560	sadness	0.531
15	OCD treatment	fear	0.682	sadness	0.402
16	OCD and gaming	sadness	0.633	fear	0.518
17	Family difficulties with OCD	fear	0.653	sadness	0.627
18	Fear of damaging possessions	sadness	0.648	disgust	0.446
19	OCD themes	fear	0.605	sadness	0.447
20	Skin picking	sadness	0.622	fear	0.512
21	Pregnancy-related OCD	fear	0.871	sadness	0.474

22	Reading and writing perfectionism	sadness	0.610	fear	0.555
23	Pedophilic OCD	sadness	0.460	optimism	0.396
24	Reassurance-seeking	optimism	0.524	sadness	0.415
25	Numerical OCD	fear	0.508	sadness	0.400
26	Music OCD and therapeutic music	sadness	0.501	joy	0.382
27	Dental OCD	fear	0.698	sadness	0.530
28	Flight anxiety	fear	0.885	sadness	0.577
29	Parenting advice for OCD child	fear	0.590	sadness	0.511
30	Alcohol and OCD	fear	0.863	sadness	0.420
31	Cannabis and OCD	fear	0.773	sadness	0.483
32	Health anxiety	fear	0.937	sadness	0.641
33	Somatic OCD	sadness	0.637	fear	0.484

Note. The primary and secondary emotion probability scores represent the probability of the specific emotion being expressed in that topic.

The three most negative topics were Topic 13 (Guilt Sensitivity and Real Event OCD), with 96% negative sentiment and a dominant emotion of sadness; Topic 32 (Health Anxiety), with 93% negative sentiment and dominant emotion fear; and Topic 12 (Food-Related OCD), with 92% negative sentiment and dominant emotion fear. Their corresponding negative emotion probability scores were sadness at 78.1%, 64.1%, and 60.8%; fear at 36.2%, 93.7%, and 73.1%; and disgust at 41.4%, 29.8%, and 31.9%, respectively.

The three most positive topics were topic 23 (*Pedophilic OCD*), with 23% positive sentiment and dominant emotion sadness; topic 31 (*Cannabis and OCD*), with 20% positive sentiment and dominant emotion fear; and topic 26 (*Music OCD and Therapeutic Music*), with 20% positive sentiment and dominant emotion sadness. Their corresponding positive emotion probability scores were joy at 33.8%, 38.2%, and 0.0%; optimism at 39.6%, 35.1%, and 0.0%; and love at 10.6%, 6.0%, and 0.0%, respectively.

The highest fear levels were found in topic 32 (*Health Anxiety*), at 93.7%, followed by topic 28 (*Flight Anxiety*), at 88.5%, and topic 21 (*Pregnancy-Related OCD*), at 87.1%. The highest disgust levels were detected in topic 14 (*Social Media Obsession*), at 52.4%; topic 20 (*Skin Picking*), at 45.1%; and topic 18 (*Fear of Damaging Possessions*), at 44.6%. The highest sadness levels were observed in topic 13 (*Guilt Sensitivity and Real Event OCD*), at 78.1%; topic 2 (*Relationship and Sexuality OCD*), at 71.7%; and topic 5 (*Existential OCD*), at 68.9%.

Discussion

This study conducted the first systematic analysis of topics, sentiments and emotions in OCD discourse on Reddit using transformer-based text mining approaches. The analysis examined 5,083 publicly available posts from r/OCD, the largest Reddit community dedicated to OCD. Analysis identified 34 distinct topics mainly revolving around specific OCD obsessions and compulsions. The posts tended to be characterized by a predominantly negative sentiment, with the majority expressing negative emotions such as fear and sadness. These findings present novel insight into the lived experiences of individuals with OCD drawn from firsthand accounts that capture both well-established and previously understudied aspects of the disorder.

Interpreting the Findings

Topic Findings

Through topic modelling, posts could be clustered into topics of varied sizes and representing a broad diversity. Several underlying thematic groupings were found. The first theme consisted of posts that addressed specific obsessions and compulsions that align with one or more well-researched OCD dimensions, such as sexuality, harm, contamination, symmetry or perfectionism, or less commonly known symptoms, such as scrupulosity or musical obsessions (Stein et al., 2019). Posts in this group could contain detailed discussions

of the obsession content, associated compulsions, their effect on daily life, requests for support or other's prior experience, and disclosure of feelings towards the symptoms. The second major theme concerned posts about treatment of OCD, such as specific treatment approaches (e.g., medication, therapy, ERP). The third theme discussed core concepts across OCD presentations, namely obsession themes, compulsions, reassurance-seeking or intrusive thoughts. Another theme addressed the relationship between substance use (e.g., cannabis, alcohol) and OCD. A fifth theme addressed the effect of OCD on familial or romantic relationships, both from the perspective of the affected person and relatives or partners who ask for advice in supporting their loved one.

Several topics emerged that have received limited attention in clinical literature so far. These topics not only expand our understanding of OCD's heterogeneous presentation but also reveal how individuals conceptualize and discuss aspects of their condition that may be underrepresented in traditional clinical assessments. The following details four particularly notable findings that offer new insights into the lived experience of OCD.

First, topic analysis revealed a specific discourse surrounding the effect of OCD on social media usage. Users describe feeling 'contaminated' by social media, frequently deleting their accounts or posts to achieve 'a clean slate', performing checking rituals, or ruminating over old messages. Indeed, research has shown that all dimensions of OCD can be affected by SM in terms of mood (Guazzini et al., 2022). For instance, James et al. (2017) coined the term *online social network obsessive-compulsive disorder* (OSN OCD) to describe maladaptive addictive behaviours motivated by a dependency on SM to fulfill one's need for belonging and social enhancement. It is characterized by intrusive obsessive thoughts about OSNs and compulsive use to soothe the corresponding anxiety. Overall, research has established that individuals with OCD symptoms are generally more vulnerable to negative effects of SM use, such as preoccupation with SM and difficulty limiting usage than non-

OCD individuals (Fontes-Perryman et al., 2022; Guazzini et al., 2022). Excessive and compulsive social media use can lead to media fatigue, which has been linked to depression, anxiety and reduced overall wellbeing (Dhir et al., 2018). However, research on the mechanism, prevalence and severity of how other dimensions of OCD affect the use of social media is limited. Posts in the current study pointed to a type of 'clean slate' approach to SM that may be indicative of the dimensions 'just right' or perfectionism OCD. Similarly, users report feelings of contamination which may point to contamination OCD. Moreover, NOCD, a major OCD treatment provider, reports that Real Event OCD, characterized by intense moral fears about past events, may lead individuals to delete their posts or SM account due to intense fear or being 'cancelled' for past online behaviour (Migala, 2023; Siev et al., 2011). Thus, more research is needed to determine the different ways in which SM and OCD may interact, for instance by using a longitudinal approach to track SM usage behaviour in diagnosed OCD samples, combined with dimension-specific SM behaviour scales.

Pregnancy-related obsessions emerged as another notable topic. In posts clustered in this topic, users detailed fears of harm towards the child, paternity doubts, and birth-related intrusive thoughts. These symptoms remain underrecognized in OCD research and treatment despite documented evidence of increased perinatal and postpartum OCD prevalence (Hudepohl et al., 2022). The discussion of these symptoms in an anonymous Reddit forum may also point to disclosure barriers in clinical settings, where stigma and child protection concerns are known to affect help-seeking behaviors (Hudepohl et al., 2022). This suggests that online communities may serve as alternative spaces for disclosure, whereas traditional clinical settings feel unsafe. The anonymity of Reddit may reduce fears of judgement or being reported, allowing parents to seek support for highly stigmatized symptoms they could not discuss elsewhere. When left untreated, perinatal OCD carries documented risks for the birth giver's psychological wellbeing, fetal and infant development, and family functioning

(Collardeau et al., 2019). These findings underscore the need for improved clinical recognition, strategies for clinicians to approach these symptoms empathetically and accessible treatment pathways for this vulnerable population.

Other users in this topic described persistent fears of unwanted pregnancy and birth control failure, compulsive pregnancy testing, and hyperfocus on bodily symptoms that might indicate pregnancy. While this aspect of pregnancy-related OCD is underrecognized in clinical research, NOCD has described the phenomenon as magical impregnation fears, concerns about pregnancy, often despite minimal or no sexual contact (Samson, 2022). This may relate to magical thinking, a well-documented cognitive pattern in OCD characterized by exaggerated expectations of negative events despite lack of reasonable evidence (Yorulmaz et al., 2011). However, further research is needed to understand the relationship between pregnancy-related obsessions and magical thinking patterns in OCD.

Another salient topic was found at the intersection of family and OCD. Some users reported feeling misunderstood by their parents and siblings, being labelled ‘disgusting’, ‘too sensitive’, or ‘attention-seeking’. Other users in this topic, on the other hand, reported about their relatives’ OCD symptoms affecting them. Another subset of posts described the exacerbation of contamination obsessions in specific family settings. Overall, users describe distress caused by stigmatization from close relatives. These reports align with previous findings that both children and adults with OCD and their relatives found family functioning to be significantly impaired (Renshaw et al., 2005). Family responses were found to range from highly antagonistic, even punitive, to enabling (Renshaw et al., 2005). The occurrence of these discussions on Reddit indicates that strategies and education for family members of OCD sufferers may be underutilized. Accordingly, family involvement should be prioritized in OCD treatment protocols.

Lastly, a topic consisting of posts discussing sleep emerged, sharing difficulties falling asleep due to OCD symptoms and having dreams adjacent to their obsessions and intrusive thoughts during wake phases. Sleep turning up as a distinct topic is consistent with a previous text mining study on Reddit mental health communities, which found sleep-related problems to be discussed as a common theme across anxiety, depression and PTSD subreddits (Park et al., 2018). This suggests that sleep issues are a transdiagnostic problem across mental health conditions.

Sentiment and Emotion Findings

The sentiment in discussions on r/OCD was overwhelmingly negative, with 31 of 34 topics found to contain posts with predominantly negative sentiment, consistent with the well-documented emotional burden and distressing nature of OCD symptoms (Reddy et al., 2017). In line with that, sadness, and fear were the most prevalent emotions found across all topics. This aligns with two of the most common comorbidities of OCD treatment-seeking patients, namely depression and anxiety, respectively (Reddy et al., 2017). In addition, disgust, concurrent with the often deviant and disturbing content of obsessions (e.g., contamination or sexual dimension), and pessimism were also frequently assigned to the posts. Unexpectedly, the highly stigmatized topic of Pedophilic OCD (POCD) contained the highest levels of positive emotion. This is not in line with research, which characterizes Pedophilic OCD as ego-dystonic, often misunderstood and followed by grave social or criminal consequences (Bonagura et al., 2022). Users in this topic often preferred to refer to posts using only the acronym ‘POCD’ and requests for private messages, either because they considered the content too distressing for others to see, or because even anonymity did not suffice in relieving the shame attached to this subject. Hence, the positive sentiment may reflect the supportive nature of the comments, not the descriptions of POCD. In addition, the

emotion scores may not just reflect POCD, as this topic contains posts that, based on their content, are not clearly about this type of obsession.

Strengths, Limitations & Future Research

There are a number of potential strengths and limitations concerning the results of this study. The process of this study relied on a machine-driven text analysis approach to replace manual analysis of r/OCD posts, allowing for a much larger scale and thus more representative and replicable results. These methods also allowed for relatively objective, data-driven insights that could otherwise not have been produced. However, models such as *BERT*, which *BERTopic* and *roBERTa-base* are based on, are themselves trained on large datasets that may also introduce bias, which could have affected the results (Bommasani et al., 2021; Mei et al., 2023). For example, language about mental illness is more likely to be classified as negative in sentiment (Mei et al., 2023). Second, BERT-based models are especially flexible in their application, as every pipeline parameter can be adjusted for optimal results, constituting a major strength that enabled the generation of meaningful and variable topics, sentiments and emotions. However, every alteration also introduces drawbacks, such as the reduction of topics, which presents the risk of producing confounded and less meaningful, convoluted topics. Thus, a balance must be found between efficiency vs complexity, tailored analysis vs replicability to achieve quality, richness, and a varied range of insights. For instance, this study was able to generate many meaningful, variable and coherent topics, however, more than a third of all posts were deemed outliers. This could be addressed in future research by further exploring how transformers could be best adjusted to the data.

The data scraping using PRAW was limited to a span of two months due to time and resource constraints and the API limit of 1000 posts. This study exclusively scraped *Hot* posts to capture the most recent and organic discussions on r/OCD. Instead, future research could

use *Top* posts, which can be scraped historically, depending on the current terms of Reddit or be scraped over a longer time span to capture seasonal or cyclical patterns of topics, sentiments and emotions.

Twitter-RoBERTa-base, the base model for both sentiment analysis and emotion recognition, was trained primarily on short-form Twitter data, while Reddit data sequences are often longer and more complex. No readily available transformer models fine-tuned on Reddit data for sentiment and emotion analysis could be identified. Future work may address this disparity by fine-tuning on Reddit data, to better capture Reddit's complexity. In addition, the emotion categories of RoBERTa-base-emotion were adapted from the SemEval-2018 Task 1 conducted by Mohammad et al. (2018), whereas future research may include emotion categories that are tailored to OCD.

The results of this study are subject to bias in terms of demographic selection. This study represents the OCD experiences of those inclined to seek support on r/OCD. A majority of Reddit users are estimated to be aged 18-34 (58%) and male (57%) (Reddit, 2020). Of those individuals, male users and users with stronger personalities, who are highly extraverted and open, and those of higher socioeconomic status are most likely to post and comment and thus may be overrepresented by these findings (Brown et al., 2018; Hargittai, 2018; Proferes et al., 2021). Moreover, members of r/OCD are not required to have an OCD diagnosis, therefore there is no clinical verification of OCD status. Accordingly, the clinical applicability of the present study may be limited. Despite these limitations, the present study can be seen as a first step towards investigating discussions about OCD among individuals with OCD on r/OCD.

Implications

This study demonstrates that transformer-based text mining presents a scalable method able to produce meaningful results from online discussions with minimal

preprocessing. The results serve as evidence that this approach can identify topics spanning well-researched and lesser-known aspects of OCD, even potentially identifying unmet needs of individuals with OCD. Accordingly, this study presents an advancement into mental health research by presenting a methodological framework that is both efficient and effective in producing insight from publicly accessible data.

The findings provide several novel insights into facets of the heterogeneous lived experience of OCD from firsthand accounts. In addition, the emotion and sentiment findings of this study clearly underlined the distressing nature of OCD, emphasizing the importance of both the present and future research. These insights have the potential to inform both clinical practice, such as treatment protocols and diagnostic scales, and OCD research by providing perspectives that have thus far been underrepresented in clinical and research settings. This study can be seen as a first step in the utilization of Reddit online forums such as r/OCD for including patient perspectives in an unobtrusive and organic way by drawing from discussions among those affected by OCD in a way that may not be possible in traditional research settings.

Opportunities for further insight are vast, including analyzing other OCD forums, or other SM platforms, such as picture- and video-based platforms. By bridging the gap between clinical knowledge and patient experiences through innovative text mining approaches, this study illustrates how the digital age offers new pathways to understanding mental health conditions. As individuals with OCD continue to seek connection and support in online spaces, systematic analysis of these discussions can provide valuable insights into patient experiences. These findings underscore the importance of incorporating patient perspectives from digital platforms into OCD research and clinical practice. On a theoretical level, *BERTopic* and *roBERTa-base-sentiment* appeared to be effective tools for investigating such discussions and drawing valuable insight about occurring topics, sentiments and emotions.

Although the generalizability and replicability of the current results must be established by future research, further text mining investigations on the patient perspectives of OCD can build on the approach, design, and available Python code used in this study.

Conclusion

The aim of the present study was to investigate discussions about OCD on r/OCD by analysing topics, sentiments and emotions through contemporary transformer-based methods. The present study established that most topics revolved around known obsessions and compulsions attributable to well-researched dimensions of OCD. Other topics concerned symptoms, treatment, and social effects of OCD. The sentiment on r/OCD was overwhelmingly negative and dominant emotions were fear and sadness for the majority of topics. These results signify a contribution to OCD research from patient perspective and demonstrate the effectiveness of BERTopic and roBERTa for OCD research on publicly available discussions.

References

Albalawi, R., Yeap, T. H., & Benyoucef, M. (2020). Using Topic Modeling Methods for Short-Text Data: A Comparative Analysis. *Frontiers in Artificial Intelligence*, *3*. https://doi.org/10.3389/frai.2020.00042

Allaoui, M., Kherfi, M. L., & Cheriet, A. (2020). Considerably improving clustering algorithms using UMAP dimensionality reduction technique: A comparative study. In A. Elmoataz, S. A. Rabah, D. Mammass, & O. Lezoray (Eds.), Image and Signal Processing (Lecture Notes in Computer Science, Vol. 12119, pp. 317–325). Springer. https://doi.org/10.1007/978-3-030-51935-3_34

American Psychiatric Association. (2022). Diagnostic and statistical manual of mental disorders (5th ed., text rev.). https://doi.org/10.1176/appi.books.9780890425787

Al-Haider, M. F., Qamar, A. M., Alkahtani, H. S., & Ahmad, H. F. (2024). Classification of obsessive-compulsive disorder symptoms in Arabic tweets using machine learning and word embedding techniques. *Journal of Advances in Information Technology, 15*(7), 798–811. https://doi.org/10.12720/jait.15.7.798-811

Antons, D., Grünwald, E., Cichy, P., & Salge, T. O. (2020). The application of text mining methods in innovation research: current state, evolution patterns, and development priorities. *R&D Management*, *50*(3), 329-351. https://doi.org/10.1111/radm.12408

Barbieri, F., Camacho-Collados, J., Espinosa Anke, L., & Neves, L. (2020). *TweetEval: Unified benchmark and comparative evaluation for tweet classification*. In Findings of the Association for Computational Linguistics: EMNLP 2020 (pp. 1644–1650). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.148

Bathje, G., & Pryor, J. (2011). The relationships of public and self-stigma to seeking mental health services. *Journal of Mental Health Counseling*, *33*(2), 161-176. https://doi.org/10.17744/mehc.33.2.g632039274160411

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77-84. http://doi.acm.org/10.1145/2133806.2133826

Boettcher, N. (2021). Studies of Depression and Anxiety Using Reddit as a Data Source: Scoping Review. *JMIR Mental Health*, *8*(11), e29487. https://doi.org/10.2196/29487

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R. B., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N. S., Chen, A. S., Creel, K., Davis, J. Q., Demszky, D., Donahue, C., … Liang, P. (2021). *On the opportunities and risks of foundation models*. arXiv. https://doi.org/10.48550/arXiv.2108.07258

Bonagura, A., Abrams, D., & Teller, J. (2022). Diagnostic Differential Between Pedophilic-OCD and Pedophilic Disorder: An Illustration with Two Vignettes. *Archives of Sexual Behavior*, *51*(4), 2359-2368. https://doi.org/10.1007/s10508-021-02273-5

Brown, D. K., Ng, Y. M. M., Riedl, M. J., & Lacasa-Mas, I. (2018). Reddit's veil of anonymity: Predictors of engagement and participation in media environments with hostile reputations. *Social Media+ Society*, *4*(4), 2056305118810216. https://doi.org/10.1177/2056305118810216

Camacho-Collados, J., Rezaee, K., Riahi, T., Ushio, A., Loureiro, D., Antypas, D., Boisson, J., Espinosa Anke, L., Liu, F., & Martínez-Cámara, E. (2022). *TweetNLP: Cutting-edge natural language processing for social media*. In W. Che & E. Shutova (Eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 38–49). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-demos.5

Chowdhary, K. R. (2020). *Natural language processing*. In *Fundamentals of Artificial Intelligence* (pp. 603–649). Springer. https://doi.org/10.1007/978-81-322-3972-7_19

Cipolla, S., Catapano, P., Pascolo, S., Luciano, M., Sampogna, G., Perris, F., ... & Catapano, F. (2024). Does duration of untreated illness impact long-term outcome in obsessive-compulsive disorder?. *European Psychiatry*, *67*(S1), S355-S356. https://doi.org/10.1192/j.eurpsy.2024.733

Collardeau, F., Corbyn, B., Abramowitz, J., & Fairbrother, N. (2019). Maternal unwanted and intrusive thoughts of infant-related harm, obsessive-compulsive disorder and depression in the perinatal period: Study protocol. *BMC Psychiatry, 19*(1), 94. https://doi.org/10.1186/s12888-019-2077-4

De Choudhury, M., & De, S. (2014). Mental Health Discourse on reddit: Self-Disclosure, Social Support, and Anonymity. *Proceedings of the International AAAI Conference on Web and Social Media*, *8*(1), 71-80. https://doi.org/10.1609/icwsm.v8i1.14526

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of deep bidirectional transformers for language understanding* [arXiv preprint arXiv:1810.04805]. arXiv. https://doi.org/10.48550/arXiv.1810.04805

Dhir, A., Yossatorn, Y., Kaur, P., & Chen, S. (2018). Online social media fatigue and psychological wellbeing—A study of compulsive use, fear of missing out, fatigue, anxiety and depression. *International Journal of Information Management*, *40*, 141-152. https://doi.org/10.1016/j.ijinfomgt.2018.01.012

Egger, R., & Yu, J. (2022). A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts. *Frontiers in sociology*, *7*, 886498. https://doi.org/10.3389/fsoc.2022.886498

Fennell, D., & Boyd, M. (2014). Obsessive-compulsive disorder in the media. *Deviant Behavior*, *35*(9), 669-686. https://doi.org/10.1080/01639625.2013.872526

Fennell, D., & Liberato, A. S. Q. (2007). Learning to Live with OCD: Labeling, the Self, and Stigma. *Deviant Behavior*, *28*(4), 305-331. https://doi.org/10.1080/01639620701233274

Fontes-Perryman, E., & Spina, R. (2022). Fear of missing out and compulsive social media use as mediators between OCD symptoms and social media fatigue. *Psychology of Popular Media*, *11*(2), 173. https://psycnet.apa.org/doi/10.1037/ppm0000356

Giordani, R. C. F., & Silva, F. S. (2021). The ethereal bodies of pro-Ana blogs: Emotional communities and spaces of sociability on the web. *Ciência & Saúde Coletiva*, *26*, 5293–5301. https://doi.org/10.1590/1413-812320212611.3.34522019

Glazier, K., Swing, M., & McGinn, L. K. (2015). Half of obsessive-compulsive disorder cases misdiagnosed: vignette-based survey of primary care physicians. *The Journal of Clinical Psychiatry*, *76*(6), 7995. http://dx.doi.org/10.4088/JCP.14m09110

Gomes, M., Pérez, M. P., Castro, I., Moreira, P., Ribeiro, S., Mota, N. B., & Morgado, P. (2023). Speech graph analysis in obsessive-compulsive disorder: The relevance of dream reports. *Journal of Psychiatric Research*, *161*, 358-363. https://doi.org/10.1016/j.jpsychires.2023.03.035

Glez-Peña, D., Lourenço, A., López-Fernández, H., Reboiro-Jato, M., & Fdez-Riverola, F. (2014). Web scraping technologies in an API world. Briefings in Bioinformatics, 15(5), 788-797. https://doi.org/10.1093/bib/bbt026

Goodman, W. K. (2014). Obsessive Compulsive and Related Disorders, an Issue of Psychiatric Clinics of North America. https://doi-org.ezproxy2.utwente.nl/10.1016/j.psc.2014.06.004

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. [arXiv preprint arXiv:2203.05794]. *arXiv.* https://doi.org/10.48550/arXiv.2203.05794

Guazzini, A., Gursesli, M. C., Serritella, E., Tani, M., & Duradoni, M. (2022). Obsessive-compulsive disorder (OCD) types and social media: are social media important and impactful for OCD people?. *European journal of investigation in health, psychology and education*, *12*(8), 1108-1120. https://doi.org/10.3390/ejihpe12080078

r/OCD. (2025, June 27). Reddit. https://www.reddit.com/r/OCD/

Reddit. (n.d.). *r/OCD* [Online community]. Reddit. Retrieved June 16, 2025, from https://www.reddit.com/r/OCD/

Hargittai, E. (2018). Potential Biases in Big Data: Omitted Voices on Social Media. *Social Science Computer Review*, *38*(1), 10-24. https://doi.org/10.1177/0894439318788322

Harrison, P. J., Cowen, P., Burns, T., & Fazel, M. (2017). *Shorter Oxford textbook of psychiatry* (7th ed.). Oxford university press. https://books.google.de/books?id=RbQ1DwAAQBAJ

Hassani, H., Beneki, C., Unger, S., Mazinani, M. T., & Yeganegi, M. R. (2020). Text Mining in Big Data Analytics. *Big Data and Cognitive Computing*, *4* (1), 1. https://doi.org/10.3390/bdcc4010001

Hirschtritt, M. E., Bloch, M. H., & Mathews, C. A. (2017). Obsessive-compulsive disorder: advances in diagnosis and treatment. *Jama*, *317*(13), 1358-1367. https://doi.org/10.1001/jama.2017.2200

Homonoff, Z., & Sciutto, M. J. (2019). The effects of obsession type and diagnostic label on OCD stigma. *Journal of obsessive-compulsive and related disorders*, *23*, 100484. https://doi.org/10.1016/j.jocrd.2019.100484

Hudepohl, N., MacLean, J. V., & Osborne, L. M. (2022). Perinatal Obsessive–Compulsive Disorder: Epidemiology, Phenomenology, Etiology, and Treatment. Current Psychiatry Reports, 24(4), 229-237. https://doi.org/10.1007/s11920-022-01333-4

James, T. L., Lowry, P. B., Wallace, L., & Warkentin, M. (2017). The effect of belongingness on obsessive-compulsive disorder in the use of online social networks. *Journal of Management Information Systems*, *34*(2), 560-596. https://dx.doi.org/10.1080/07421222.2017.1334496

Jurafsky, D., & Martin, J. H. (2025). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (3rd ed.). [Online draft]. Retrieved from https://web.stanford.edu/~jurafsky/slp3/

Kemp, S. (2024, January 31). *Digital 2024: Global overview report*. *DataReportal*. https://datareportal.com/reports/digital-2024-global-overview-report

Keyes, C., Nolte, L., & Williams, T. I. (2017). The battle of living with obsessive compulsive disorder: a qualitative study of young people's experiences. *Child and Adolescent Mental Health*, *23*(3), 177-184. https://doi.org/10.1111/camh.12216

Kim, S., Cha, J., Kim, D., & Park, E. (2023). Understanding Mental Health Issues in Different Subdomains of Social Networking Services: Computational Analysis of Text-Based Reddit Posts. *Journal of medical Internet research*, *25*, e49074. https://doi.org/10.2196/49074

Kherwa, P., & Bansal, P. (2018). Topic Modeling: A Comprehensive Review. *ICST Transactions on Scalable Information Systems*, 159623. https://doi.org/10.4108/eai.13-7-2018.159623

Kohler, K. C., Coetzee, B. J. S., & Lochner, C. (2018). Living with obsessive-compulsive disorder (OCD): a South African narrative. *International Journal of Mental Health Systems*, *12*, 1-11. https://doi.org/10.1186/s13033-018-0253-8

Lau, J. H., Newman, D., & Baldwin, T. (2014). Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 530–539). Association for Computational Linguistics. https://doi.org/10.3115/v1/E14-1056

Lee, S. L., Park, M. S. A., & Tam, C. L. (2015). The relationship between Facebook attachment and obsessive-compulsive disorder severity. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, *9*(2). https://doi.org/10.5817/CP2015-2-6

Liu, Y., Shan, Y., Sun, S., Ji, M., Zhou, S., You, Y., Liu, H., & Shen, Y. (2024). Topic modeling and content analysis of people's anxiety-related concerns raised on a computer-mediated health platform. *Scientific Reports*, *14*(1). https://doi.org/10.1038/s41598-024-79164-x

Lundblade, K. M. (2023). Sorting Things Out: Critically Assessing the Impact of Reddit's Post Sorting Algorithms on Qualitative Analysis Methods. In *Proceedings of the 41st ACM International Conference on Design of Communication* (pp. 112-118). https://doi-org.ezproxy2.utwente.nl/10.1145/3615335.3623021

McCormack, A., & Coulson, N. S. (2009). Individuals with eating disorders and the use of online support groups as a form of social support. Cyberpsychology. *Journal of Psychosocial Research on Cyberspace*, *3*(2). https://doi.org/10.1097/ncn.0b013e3181c04b06

Medvedev, A. N., Lambiotte, R., & Delvenne, J.-C. (2018). The anatomy of Reddit: An overview of academic research [arXiv preprint arXiv:1810.10881]. *arXiv*. https://doi.org/10.48550/arXiv.1810.10881

Mei, K. X., Fereidooni, S., & Caliskan, A. (2023). Bias against 93 stigmatized groups in masked language models and downstream sentiment classification tasks. In *2023 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1-20). ACM. https://doi.org/10.1145/3593013.3594109

Migala, J. (2023, September 18). *I'm afraid of people finding my old social media posts. What can I do?* NOCD. https://www.treatmyocd.com/what-is-ocd/common-fears/im-afraid-of-people-finding-my-old-social-media-posts-what-can-i-do

Mohammad, S., Bravo-Marquez, F., Salameh, M., & Kiritchenko, S. (2018). SemEval-2018 Task 1: Affect in Tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation* (pp. 1-17). Association for Computational Linguistics. https://doi.org/10.18653/v1/S18-1001

Nandwani, P., & Verma, R. (2021). A review on sentiment analysis and emotion detection from text. Social Network Analysis and Mining, 11(1). https://doi.org/10.1007/s13278-021-00776-6

Park, A., Conway, M., & Chen, A. T. (2018). Examining thematic similarity, difference, and membership in three online mental health communities from reddit: A text mining and visualization approach. *Computers in Human Behavior*, *78*, 98-112. https://doi.org/10.1016/j.chb.2017.09.001

Patel, S. R., Galfavy, H., Kimeldorf, M. B., Dixon, L. B., & Simpson, H. B. (2017). Patient Preferences and Acceptability of Evidence-Based and Novel Treatments for Obsessive-Compulsive Disorder. *Psychiatric Services*, *68*(3), 250-257. https://doi.org/10.1176/appi.ps.201600092

Pavelko, R. L., & Myrick, J. G. (2015). That's so OCD: The effects of disease trivialization via social media on user perceptions and impression formation. *Computers in Human Behavior*, *49*, 251-258. https://doi.org/10.1016/j.chb.2015.02.061

Perez, M. I., Limon, D. L., Candelari, A. E., Cepeda, S. L., Ramirez, A. C., Guzick, A. G., ... & Storch, E. A. (2022). Obsessive-compulsive disorder misdiagnosis among mental healthcare providers in Latin America. *Journal of obsessive-compulsive and related disorders*, *32*, 100693. https://doi-org.ezproxy2.utwente.nl/10.1016/j.jocrd.2021.100693

Plutchik, R. (1980). A general psychoevolutionary theory of emotion. In R. Plutchik & H. Kellerman (Eds.), *Theories of emotion* (pp. 3-33). Academic Press. https://doi.org/10.1016/B978-0-12-558701-3.50007-7

Ponzini, G. T., & Steinman, S. A. (2022). A systematic review of public stigma attributes and obsessive–compulsive disorder symptom subtypes. *Stigma and Health*, *7*(1), 14. https://doi.org/10.1037/sah0000310

Proferes, N., Jones, N., Gilbert, S., Fiesler, C., & Zimmer, M. (2021). Studying reddit: A systematic overview of disciplines, approaches, methods, and ethics. *Social Media+ Society*, *7*(2), 20563051211019004.https://doi.org/10.1177/20563051211019004

Reddit. (2020, June). *Audience* [Web page]. *Reddit Inc.* https://web.archive.org/web/20210117184818/https://www.redditinc.com/advertising/audience#tab2

Reddit. (2024, November 6). *What are communities or subreddits? Reddit Help*. https://support.reddithelp.com/hc/en-us/articles/204533569-What-are-communities-or-subreddits

Reddit, Inc. (2025, May 1). *Quarterly report on Form 10-Q for the quarter ended March 31, 2025*. Retrieved June 16, 2025, from https://s203.q4cdn.com/380862485/files/doc_financials/2025/q1/80fd347c-6d90-4264-8bac-b8ff9f56cb02.pdf

Reddy, Y. J., Sundar, A. S., Narayanaswamy, J. C., & Math, S. B. (2017). Clinical practice guidelines for obsessive-compulsive disorder. *Indian journal of psychiatry*, *59*(Suppl 1), S74. https://doi.org/10.3390/ejihpe12080078

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3982–3992. https://doi.org/10.18653/v1/D19-1410

Renshaw, K. D., Steketee, G., & Chambless, D. L. (2005). Involving Family Members in the Treatment of OCD. *Cognitive Behaviour Therapy*, *34*(3), 164-175. https://doi.org/10.1080/16506070510043732

Rezapour, M. (2024). Emotion Detection with Transformers: A Comparative Study. *arXiv preprint arXiv:2403.15454*. https://doi.org/10.48550/arXiv.2403.15454

Röder, M., Both, A., & Hinneburg, A. (2015, February). Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining* (pp. 399-408). https://doi.org/10.1145/2684822.2685324

r/OCD. (n.d.). *About r/OCD* [Online forum page]. *Reddit*. Retrieved February 3, 2025, from https://www.reddit.com/r/OCD/about/

r/OCD. (2023, November 25). *Reassurance seeking and providing: Rules of this subreddit.* [Online forum post]. *Reddit*. https://www.reddit.com/r/OCD/comments/17xi2ao/reassurance_seeking_and_providing_rules_of_this/

Salihefendic, A. (2015, December 8). *How Reddit ranking algorithms work*. *Hacking and Gonzo*. Medium. https://medium.com/hacking-and-gonzo/how-reddit-ranking-algorithms-work-ef111e33d0d9

Samson, A. (2022, November 11). *Fear of magical impregnation OCD*. TreatMyOCD. https://www.treatmyocd.com/what-is-ocd/common-fears/fear-of-magical-impregnation-ocd

Schofield, C. A., & Ponzini, G. T. (2020). The Skidmore Anxiety Stigma Scale (SASS): A covert and brief self-report measure. *Journal of anxiety disorders*, *74*, 102259. https://doi.org/10.1016/j.janxdis.2020.102259

Sidani, J. E., Shensa, A., Hoffman, B., Hanmer, J., & Primack, B. A. (2016). The Association between Social Media Use and Eating Concerns among US Young Adults. *Journal of the Academy of Nutrition and Dietetics*, *116*(9), 1465–1472. https://doi.org/10.1016/j.jand.2016.03.021

Siev, J., Baer, L., & Minichiello, W. E. (2011). *Obsessive-compulsive disorder with predominantly scrupulous symptoms: Clinical and religious characteristics*. *Journal of Clinical Psychology, 67*(12), 1188–1196. https://doi.org/10.1002/jclp.20843

Sit, M., Elliott, S. A., Wright, K. S., Scott, S. D., & Hartling, L. (2022). Youth Mental Health Help-Seeking Information Needs and Experiences: A Thematic Analysis of Reddit Posts. *Youth &Amp; Society*, *56*(1), 24-41. https://doi.org/10.1177/0044118x221129642

Sravanti, L., Kommu, J. V. S., Girimaji, S. C., & Seshadri, S. (2022). Lived experiences of children and adolescents with obsessive–compulsive disorder: interpretative phenomenological analysis. *Child and Adolescent Psychiatry and Mental Health*, *16*(1). https://doi.org/10.1186/s13034-022-00478-7

Steinberg, D. S., & Wetterneck, C. T. (2017). OCD taboo thoughts and stigmatizing attitudes in clinicians. *Community mental health journal*, *53*, 275-280. https://doi.org/10.1007/s10597-016-0055-x

Subramaniam, M., Soh, P., Ong, C., Esmond Seow, L. S., Picco, L., Vaingankar, J. A., & Chong, S. A. (2014). Patient-reported outcomes in obsessive-compulsive disorder. Dialogues in Clinical Neuroscience, 16(2), 239-254. https://doi.org/10.31887/dcns.2014.16.2/

Sukumar, R., Sharma, K., Rajan, T. S., Deshpande, Y. D., Agarwal, A., & Nagpal, M. (2024, June). Exploring the Benefits of Text Mining for Automated Information Discovery. In *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)* (pp. 1-6). IEEE. https://doi-org.ezproxy2.utwente.nl/10.1109/ICCCNT61001.2024.10724762

Tan, Y., Rehm, I., Stevenson, J., & De Foe, A. (2021). Social media peer support groups for obsessive-compulsive and related disorders: Understanding the predictors of negative experiences. *Journal of Affective Disorders*, *281*, 661–672. https://doi.org/10.1016/j.jad.2020.11.094

Van Schalkwyk, G. I., Bhalla, I. P., Griepp, M., Kelmendi, B., Davidson, L., & Pittenger, C. (2016). Toward understanding the heterogeneity in obsessive-compulsive disorder:

Evidence from narratives in adult patients. *Australian & New Zealand Journal of Psychiatry*, *50*(1), 74-81. https://doi.org/10.1177/0004867415579919

Wang, Y., Qu, W., & Ye, X. (2024, November 7). *Selecting between BERT and GPT for text classification in political science research* [Preprint]. *arXiv*. https://doi.org/10.48550/arXiv.2411.05050

Wheaton, M. G., Sternberg, L., McFarlane, K., & Sarda, A. (2016). Self-concealment in obsessive-compulsive disorder: Associations with symptom dimensions, help seeking attitudes, and treatment expectancy. *Journal of Obsessive-Compulsive and Related Disorders*, *11*, 43-48. https://doi.org/10.1016/j.jocrd.2016.08.002

Woods, E. E., Gantt-Howrey, A., & Pope, A. L. (2023). "I'm so #OCD": A Content Analysis of How Women Portray OCD on TikTok. *The Professional Counselor*, *13*(1), 27-38. https://doi.org/10.15241/eew.13.1.27

Yadav, A., & Vishwakarma, D. K. (2020). Sentiment analysis using deep learning architectures: A review. *Artificial Intelligence Review*, *53*(6), 4335–4385. https://doi.org/10.1007/s10462-019-09794-5

Yorulmaz, O., Inozu, M., & Gültepe, B. (2011). The role of magical thinking in Obsessive-Compulsive Disorder symptoms and cognitions in an analogue sample. *Journal of Behavior Therapy and Experimental Psychiatry*, *42*(2), 198-203. https://doi.org/10.1016/j.jbtep.2010.11.007

Zhang, J., Hamilton, W., Danescu-Niculescu-Mizil, C., Jurafsky, D., & Leskovec, J. (2017, May). Community identity and user engagement in a multi-community landscape. In *Proceedings of the international AAAI conference on web and social media* (Vol. 11, No. 1, pp. 377-386). https://doi.org/10.48550/arXiv.1705.09665

Ziegler, S., Bednasch, K., Baldofski, S., & Rummel-Kluge, C. (2021). Long durations from symptom onset to diagnosis and from diagnosis to treatment in obsessive-compulsive

disorder: A retrospective self-report study. *PLOS ONE*, *16*(12), e0261169. https://doi.org/10.1371/journal.pone.0261169

Zisler, E. M., Meule, A., Koch, S., Schennach, R., & Voderholzer, U. (2024). Duration of daily life activities in persons with and without obsessive–compulsive disorder. *Journal of Psychiatric Research*, *173*, 6-13. https://doi.org/10.1016/j.jpsychires.2024.02.052

Zong, C., Xia, R., & Zhang, J. (2021). *Text Data Mining* Springer Singapore. https://doi.org/10.1007/978-981-16-0100-2

Zhong, Q., Ding, L., Liu, J., Du, B., & Tao, D. (2023, March 2). *Can ChatGPT understand too? A comparative study on ChatGPT and fine-tuned BERT* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2302.10198

During the preparation of this work, I used ChatGPT 4.0 to debug code, the Microsoft Word built-in word correction and suggestion tool, Lean Library and Claude to search for additional references, and to check for grammatical mistakes and language clarity. After using these tools, I thoroughly reviewed and edited the content as needed, taking full responsibility for the final outcome.