Emetophobia Dataset Collection, NER, Sentiments, and Topic Analysis

Florin-Silviu Dinu
florin-silviu.dinu@s.unibuc.ro

Abstract

A dataset of emetophobia posts from Reddit was collected and processed for the purpose of identifying key factors in the informal discussions. The dataset is mostly labeled by users from a set of labels that describe the type of post. The posts were analyzed for the specific sentiment that the user had when writing and emoji usage was processed in order to check wether the specific emojis express a different sentiment than the sentiment of the text or they simply add to the user's expressivity. Named Entity Recognition with scientific and medical models was performed in order to identify the most common discussed topics and was further enhanced by using Retrieval Augmented Generation by using the posts and an LLM for identifying other common topics. Supervised and unsupervised classification models were used in order to check the separability of the text within the labeled topics. The results presented by the paper can be further used to study the general discourse around this phobia.

1 Introduction

Phobias are characterized either by an intense disproportional fear or an aversion towards a specific situation or object (National Institute of Mental Health, 2024b) and constitute a high risk of developing other anxiety disorders (McLean Hospital, 2025). Fear is the underlying mechanism and can appear in a natural state as a mechanism that humans or other animals exhibit as a response to stress (McLean Hospital, 2025). Anxiety issues the same response as fear but at a lower level and can lead to avoidance of certain situations or objects (McLean Hospital, 2025) and is determined to affect about a third of US adolescents and adults (National Institute of Mental Health, 2024a).

According to NIMH, phobias have the following characteristics (National Institute of Mental Health, 2024b):

- exhibit an irrational and excessive worry about an object or situation
- cause an avoidance behavior directed at the specific object or situation
- in case of encountering the specific object or situation, there is an observed immediate fear and intense anxiety
- in case of a period of endurance of the specific object or situation the anxiety is intense in nature for the whole period

Even if phobias can cause other types of anxiety disorders (McLean Hospital, 2025), their cause according to NIMH may be (National Institute of Mental Health, 2024b):

- A traumatic event involving the triggering object or situation
- A general sentiment of nervousness or distress in new situations in childhood
- A genetic component

Simple phobias are also called specific phobias and can vary in the nature of the triggering object or situation (National Institute of Mental Health, 2024b). The example studied in this paper is emetophobia, also known as Specific Phobia of Vomiting (SPOV) and is classified in DSM-V in the 'Specific Phobia: Other' class (Keyes et al., 2018). The SPOV is rare in community samples, demographically is prevalent more in women, individuals that are affected generally seek treatment but has an overlap with other disorders such as OCD, health anxiety and disordered eating which may make the diagnostic more difficult compared to other phobias (Keyes et al., 2018).

Given these causes and characteristics of SPOV, this paper presents a collected dataset from Reddit of individual user posts regarding the disorder.

The used subreddit is `r/emetophobia` and can be accessed through this link: https://www.reddit.com/r/emetophobia/ (Reddit community r/emetophobia, 2025). The scripts that were used to collect and analyze the data can be found here: https://github.com/fredtux/BioNLP-scripts while the dataset will not be made public on Github due to privacy concerns.

2 Related Work

Online communication may pose certain vulnerable groups, especially children and adolescents, and is prevalent throughout all demographic categories, 59.4% of the world's population using social media as part of their lives and implicitly engaging in forms of communication (Montejo-Ráez et al., 2024). Montejo-Ráez et al. identified 4 categories of disorders that are prevalent and have comorbidities and studied the algorithms used in the surveyed papers (Montejo-Ráez et al., 2024):

- **Eating disorder/Anorexia** (survey period: 2018-2020): shallow learning approaches are a constant throughout the years, with deep learning for training being used less over time but large language models started being used since 2020
- **Depression** (survey period: 2020-2023): shallow learning is a constant throughout the surveyed period, deep learning has been used throughout the period as well but in less studies, LLMs were used since 2020 with an uptick in 2021 and hybrid approaches were used only in the first three years
- **Pathological gambling** (survey period: 2020-2023): most articles that were included in the survey were published in 2022 and included a majority of deep learning methods followed by shallow learning and LLMs
- **Suicidal ideation** (survey period: 2020-2023): shallow learning is used in most cases, deep learning comes second and hybrid approaches became more prevalent in 2023

A broader survey by Zhang et al. that encompasses the 2012-2021 period shows that deep learning started being used for mental illness NLP models in 2013 and surpassed classical machine learning in 2021 (Zhang et al., 2022) which is also supported by a survey of 2010-2022 studies by Malgaroli et al. (Malgaroli et al., 2023). When it comes to data sources, we see that both Montejo-Ráez et al. and Zhang et al. find that online sources are more prevalent, with Zhang et al. discovering that 81% of the studies have a dataset from an online source (Zhang et al., 2022) and both of them finding that the most used datasets contain data from Twitter, Reddit, Weibo, Facebook and Instagram (Montejo-Ráez et al., 2024; Zhang et al., 2022) which are social media platforms while Zhang et al. also showing that various online forums comprise 4% of the social media sources (Zhang et al., 2022). This is also supported by a survey from Glaz et al. which identifies Twitter and Reddit as the 2 main social media sources of dataset for studies (Glaz et al., 2021). Regarding the ML models, Zhang et al. identifies that their features are based on POS, BoW, Linguistic Inquiry and Word Count, n-grams, TF-IDF and others which are widely used in many NLP tasks (Zhang et al., 2022) which is supported by Glaz et al. (Glaz et al., 2021) while the models are both supervised like SVM, AdaBoost, KNN, XGBoost and others and unsupervised like LDA while semi-supervised learning is used rarely used (Zhang et al., 2022).

For both detection and treatment, Mármol-Romero and Alba Maria propose a dialogue system with conversational agents (Márḿol-Romero, 2024) and implemented along other co-authors an experimental GPT-3 chatbot for Spanish speaking teenagers which shows promising results but needs to be further refined, the authors proposing a RAG in order to deal with the constant translations (Mármol-Romero et al., 2024).

3 Method

The dataset was collected from the subreddit **r/emetophobia** (Reddit community r/emetophobia, 2025), was processed for further use, had sentiment analysis performed and NER extraction and these were used as features for supervised and unsupervised topic classification.

3.1 Dataset

The dataset consists of 40 pages scraped using the Reddit API (Reddit, Inc., 2025) which were combined into one file where only the following **data** subsections were kept: title, content, date, ups, upvote_ratio, and labels, also downvotes were computed and a "None" label was added for the articles that did not have a label. For the labels only the text first label was kept and the number of articles for

each label can be found in Table 1. Furthermore, posts with emojis were processed, identifying 148 posts that contain emojis using the emoji library (Kim et al., 2025).

Label	Num Posts
Question	199
Needing support - Panic attack	151
Rant	135
Potentially Triggering	85
Does Anyone Else...?	78
Needing support: Just not feeling good	73
Needing Support - In Acute Crisis (at risk of self injury)	45
Venting - Advice wanted	42
Needing Support - N, V, D etc	40
Success!	39
Needing Support - Anxious about FP	31
Techniques, tips and tricks	19
None	16
It Happened (TW)	12
Recovery	11
Positive Reminder	5
Needing Support - Non-Emet related	4
Therapy info!	3
Interesting info/Articles	3
Needing Support - N, V, D etc NO RE-ASSURANCE	3
Venting - No advice please	1
Moderator	1
Total	996

Table 1: Number of Posts per Label

3.2 Preprocessing

For sentiment analysis the dataset was split in 2: with and without emojis. Both of them had the new lines replaced with spaces so that the text will become more legible.

For NER new lines were also replaced and emojis were removed in order to not interfere with the models. Only posts with non-empty lists of NER were kept, causing 10 more posts to be removed, bringing the total to 986.

For topic classification identified NERs were stitched together and another array with the compound sentiment score was created.

For the chat with documents part, titles and contents were concatenated and the NER data was concatenated at the end.

3.3 Sentiment Analysis

VADER is a sentiment analysis tool that outputs a continuous number on the interval $[-1, 1]$ with the interpretation that if the number is > 0.5, the sentiment is positive, if the number is < -0.5, then the sentiment is negative, otherwise the sentiment is neutral, making it a 3-class classifier (Hutto and Gilbert, 2015, 2014). The implementation can also take emojis into account (Hutto and Gilbert, 2014).

For the overall dataset, the maximum score of a post both with emojis and without is 0.9982 which is extremely positive and the minimum score is -0.9986 which is extremely negative. The mean for the processing with emojis is -0.3116 and the standard deviation is 0.7065 and without emojis the mean is -0.3004 and the standard deviation 0.7079 which shows that even if the score is neutral, it tends to be closer to negative emotions and also that it encompasses a full range of positive and negative emotions. For a paired sample t-test between the set with emojis and that without, $t = -3.156$ and $p = 0.0017$ which shows a strong correlation and that the sentiments in the dataset without emojis are moderately more positive.

For the top 3 labels with the largest numbers of posts, the data shows that:

- **"Questions"** label is neutral with a mean of -0.1914 and a standard deviation of 0.6835 with emojis and -0.1781 mean and a 0.6863 standard deviation without emojis, having no strong correlation between the two types of processing with $p = 0.07556$

- **Needing support - Panic attack** label has negative emotions with a mean of -0.6125 and a standard deviation of 0.5249 with emojis and a mean of -0.5964 and a standard deviation of 0.5355 without emojis, having a strong correlation of $p = 0.01213$ with $t = -2.5392$, emojis being used to accentuate the negative emotions

- **Rant** label has a neutral emotion with a mean of -0.4662 and a standard deviation of 0.6447 with emojis and a mean of -0.4505 and a standard deviation of 0.6484, the two types of processing not issuing a strong correlation for the sentiment analysis with $p = 0.1846$

Furthermore, a study of the context in which emojis are used and the influence of each was conducted. This shows that the top 3 emojis produced the following effects:

- 😭 is the most used emoji appearing 91 different times giving a mean of -0.724 with a standard deviation of 0.5776 when it is taken into consideration and a mean of -0.445 and a standard deviation of 0.6646 when it is not taken into consideration, have a strong correlation of $p < 0.01$ for influencing the sentiment of the posts

- 🫶 appears just 12 times in positive posts with a mean of 0.6724 and a standard deviation of 0.5362 when it is taken into consideration and a mean of 0.5995 and a standard deviation of 0.61467 when it isn't, the two occasions not being correlated, having $p = 0.253$

- ❤️ appears 10 times and has a mean of -0.2125 with a standard deviation of 0.7806 when taken into consideration and a mean of -0.2123 and a standard deviation of 0.7804 when not, the two not being strongly correlated with $p = 0.0839$

The findings prove that the subreddit is neutral but with a tendency towards negative sentiments with the most popular labels being neutral to negative and the most used emoji being negative. The only downside of the study is the lack of a clear separation between emoji use in the same text and the lack of balancing the spamming of a single emoji in a text.

3.4 NER

For NER 2 methods were used: scispacy's **en_core_sci_md** with the **UmlsEntityLinker** for the knowledge base of extracted NERs (Neumann et al., 2019) and BioBERT (Lee et al., 2020) trained for NER extraction (Raza et al., 2022). Both of them were used on texts without emojis processed as described in Section 3.2.

For scispacy the first 3 labels ranked by the number of posts as in Table 1 had the following number of NERs: "Questions" - 2581, "Needing support - Panic attack" - 2504 and "Rant" - 2616. For BioBERT NER the first 3 labels had: Questions" - 1578, "Needing support - Panic attack" - 2101 and "Rant" - 1657. Another difference is that BioBERT NER produced only 7 posts without NER while scispacy produced 3 more.

The top 5 entities produced by scispacy were: I-antigen, Anxiety, Illness (finding), Daily and Nausea. The top 5 produced by BioBERT NER were the groups: Sign_symptom, Activity, Detailed_description, Time and Medication.

3.5 RAG

In order to further check the effects of NER, a RAG with Langchain (LangChain Contributors, 2025) that takes the documents and uses them to help Mistral Nemo (Mistral AI team, 2024) retrieve the answer to the query: **"What are the top 5 most common medical or psychological ideas that the documents refer to? Make a numbered list. Respond with just the list, avoid other text."**.

There are 3 types of documents that are being passed:

- Title concatenated with content

- Title concatenated with content, concatenated with UmlsEntityLinker's canonical name and definition

- Title concatenated with content, concatenated with BioBERT's NER entity and group

The processed results are presented in Table 2 for no NER, Table 3 for scispacy and Table 4 for BioBERT NER.

NO NER
Exposure Therapy
Emetophobia
OCD
Anti-Anxiety Medication
CBT-Based Guide to Emetophobia

Table 2: Labels Extracted with NO NER

Scispacy
Panic Attacks
Nausea/Vomiting
PTSD
Zofran
Mental Suffering/Distress

Table 3: Labels Extracted with Scispacy

BioBERT
Exposure Threapy
Emetophobia
Panic Attacks
Reassurance Seeking (Harmful in Context)
Anxiety Disorder

Table 4: Labels Extracted with BioBERT NER

Correlating scispacy NER RAG with scispacy most common entities, we see **nausea** appearing in both and **anxiety** and **illness (finding)** being somewhat correlated with **panic attacks**, **PTSD** and **mental suffering/distress**. Comparing it to BioBERT NER's entity groups, we see **sign_symptom**, **activity** and **medication** being correlated with **panic attacks**, **vomiting** and **Zofran**.

Correlating BioBERT NER's RAG with scispacy, we see **illness (finding)**, **anxiety** and **nausea** correlated with **panic attacks**, **anxiety disorder** and **emetophobia**. Comparing it to BioBERT NER's groups, we see **sign_symptom** and **activity** correlated with **panic attacks**, **anxiety disorder**, and **Reassurance Seeking (Harmful in Context)**

Correlating no NER's RAG with scispacy, we see **nausea** appearing in **emetophobia** and with BioBERT's NER **sign_symptom**, **detailed_description** and **medication** appear in **emetophobia**, **OCD**, **CBT-Based Guide to Emetophobia** and **Anti-Anxiety Medication**.

Making these 3 correlation an improvement in both specificity and resemblance with the extracted NERs can be seen in adding scispacy UmlsEntityLinker's canonical name and definition to the RAG.

3.6 Topic Classification

3.6.1 Train-test split

For the topic classification, a train-test split of 0.8-0.2 was made using the following labels: 'Question', 'Needing support - Panic attack', 'Rant', 'Potentially Triggering', 'Does Anyone Else...?', 'Needing support: Just not feeling good'. Only posts with greater than 0 entities were kept, resulting in 400-101 split for scispacy's NER and 338-85 split for BioBERT's NER. For the contents, the token **<PAD>** was added to meet the 128 token limit per post for the posts that had less and a trim was made for the posts that had more. For the emotions array, only the emotions of the full text with emojis were included.

3.6.2 Unsupervised

For **DBSCAN** a skipgram model was trained with a vector size of 64, a window of 8 for scispacy and 10 for BioBERT and all words were included. The mean of each vector was taken and the compound emotion score was added at the end. DBSCAN was trained on a maximum of 300 iterations on 6 clusters. For scispacy a silhouette score of 0.6441 was computed and for BioBERT 0.6348. Both scores show a clear boundary between the classes.

For **BERTopic** (Grootendorst, 2022) sentence-transformers/all-MiniLM-L6-v2 (Reimers et al., 2022) was used and then, BERTopic was used with 1 and 2 n-grams, a minimum topic size of 2 and 6 clusters. The result for scispacy was 5 clusters plus noise with a silhouette score of 0.0346 without the noise cluster and for BioBERT there were again 5 clusters without the noise one and a silhouette score of 0.0384. This shows that the sentence transformer on named entities coupled with BERTopic doesn't render groups that are clearly delineated.

3.6.3 Supervised

For **BERTopic** (Grootendorst, 2022) supervised sentence-transformers/all-MiniLM-L6-v2 (Reimers et al., 2022) was used and it was coupled with a SVC (scikit-learn developers, 2025) with the RBF kernel and the default parameters and issued an accuracy of 0.3564 for scispacy's NER and 0.0941 for BioBERT's NER. This shows again that the sentence transformer on named entities coupled with BERTopic doesn't make good enough predictions. The confusion matrix for scispacy's NER can be seen in Figure 1 and for BioBERT's NER in Figure 2.

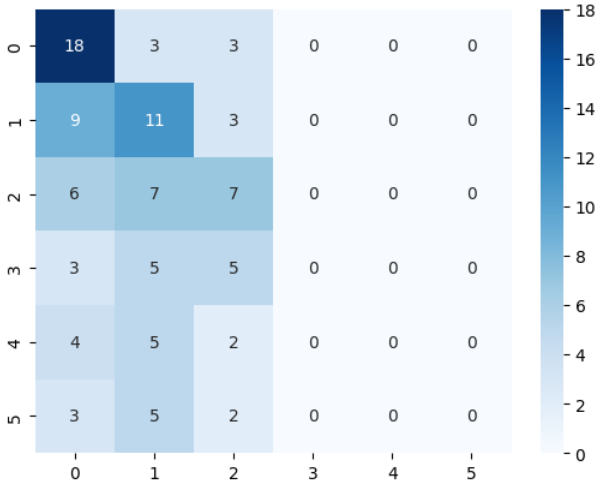

Figure 1: BERTopic supervised with scispacy NER

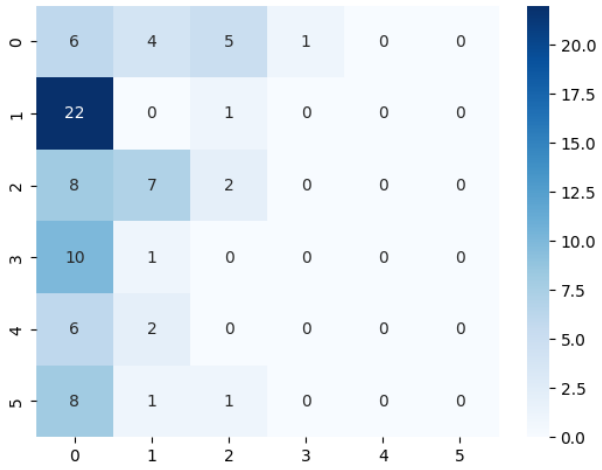

Figure 2: BERTopic supervised with BioBERT NER

For **XGBoost Classifier** a skipgram model like in Section 3.6.3 was trained and a grid search for parameters was used with an f1_macro scoring strategy. The resulting model had a learning rate of 0.1, a maximum depth of 3 and 100 estimators for scispacy and a learning rate of 0.2, a maximum depth of 3 and 200 estimators for BioBERT. An accuracy of 0.2277 for scispacy and 0.2 for BioBERT was achieved. The confusion matrix for scispacy's NER can be seen in Figure 3 and for BioBERT's NER in Figure 4.

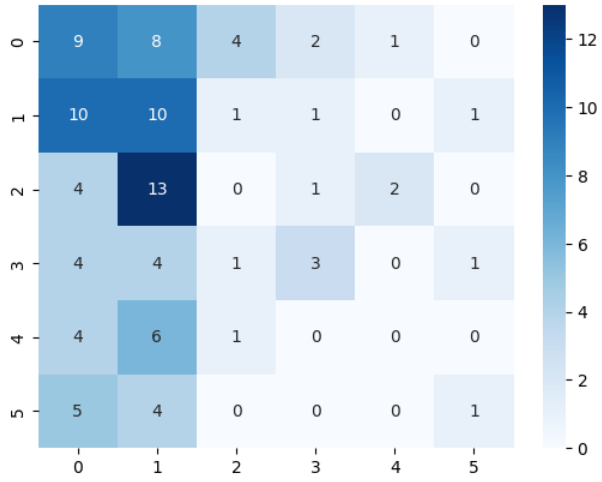

Figure 3: XGBoost Classifier with scispacy NER

None of the classifiers have a good enough result and improvements have to be considered.

4 Conclusion

The emetophobia dataset could have fallen in line with other SOTA datasets if there was more data. Even so the dataset looks promising for the following reasons:

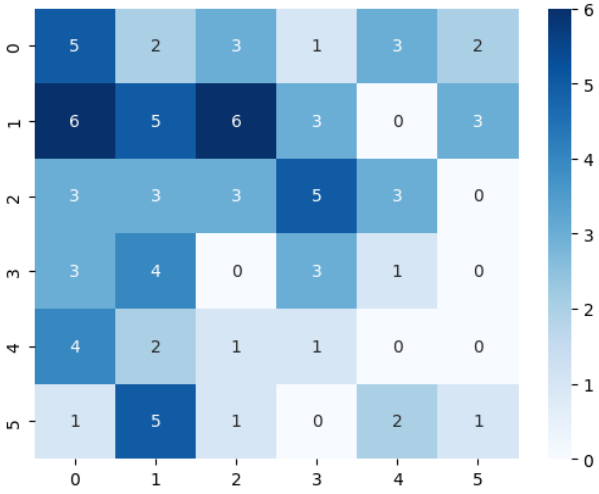

Figure 4: XGBoost Classifier with BioBERT NER

- Sentiments can be clearly delineated
- In-group emoji use can be easily studied
- An improvement in RAG applications can be found if using scispacy UmlsEntityLinker's canonical names and definitions
- Usual NER models fare well on the dataset
- Labels can be delineated with a good enough silhouette score through DBSCAN but the rest of the models may lack data or may not fare well with the tried preprocessing methods

Ethical Statement

The dataset is not made public on Github because it might be used to target individuals. Also, this project has the purpose of finding insights into the public informal SPOV discourse in order to better understand the people suffering from this disorder and as a basis for new diagnosis methods but not in order to target, harass or discriminate against them.

References

Aziliz Le Glaz, Yannis Haralambous, Deok-Hee Kim-Dufor, Philippe Lenca, Romain Billot, Taylor C. Ryan, Jonathan Marsh, Jordan DeVylder, Michel Walter, Sofian Berrouiguet, and Christophe Lemey. 2021. Machine learning and natural language processing in mental health: Systematic review. *Journal of Medical Internet Research*, 23(5):e15708. Originally published in the Journal of Medical Internet Research, 04.05.2021.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

C. J. Hutto and E. E. Gilbert. 2014. Vader sentiment analysis. GitHub repository. Latest public release "0.5" on Nov 17, 2014; MIT license; Accessed 16 June 2025.

C.J. Hutto and Eric Gilbert. 2015. Vader: A parsimonious rule-based model for sentiment analysis of social media text.

Alexandra Keyes, Helen R. Gilpin, and David Veale. 2018. Phenomenology, epidemiology, co-morbidity and treatment of a specific phobia of vomiting: A systematic review of an understudied disorder. *Clinical Psychology Review*, 60:15–31.

Taehoon Kim, Kevin Wurster, and Tahir Jalilov. 2025. emoji: Emoji terminal output for python. GitHub repository. Latest release v2.14.1; Accessed 16 June 2025.

LangChain Contributors. 2025. langchain: Framework for building llm-powered applications. GitHub repository and PyPI package. Latest PyPI release v0.3.25 (May 2, 2025); Latest GitHub Core release v0.3.65 (June 10, 2025); MIT license; Accessed 17 June 2025.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240. Published by Oxford University Press. Pre-trained model available at: https://github.com/naver/biobert-pretrained; fine-tuning code: https://github.com/dmis-lab/biobert.

Matteo Malgaroli, Thomas D. Hull, James M. Zech, and Tim Althoff. 2023. Natural language processing for mental health interventions: a systematic review and research framework. *Translational Psychiatry*, 13(1):309.

Alba María Mármol-Romero. 2024. Natural language processing in the detection and treatment of mental health issues. In *Doctoral Symposium on Natural Language Processing*, volume 3797, Valladolid, Spain. CEUR Workshop Proceedings. CC BY 4.0.

McLean Hospital. 2025. Understanding fear, anxiety, and phobias. `https://www.mcleanhospital.org/essential/fear-phobias`. Accessed 16 June 2025.

Mistral AI team. 2024. Mistral nemo. Web page, Mistral AI. Released under Apache 2.0 license; Accessed 16 June 2025.

Arturo Montejo-Ráez, M. Dolores Molina-González, Salud María Jiménez-Zafra, Miguel Ángel García-Cumbreras, and Luis Joaquín García-López. 2024. A survey on detecting mental disorders with natural language processing: Literature review, trends and challenges. *Computer Science Review*, 53:100654.

Alba María Mármol-Romero, Manuel García-Vega, Miguel Ángel García-Cumbreras, and Arturo Montejo-Ráez. 2024. An empathic gpt-based chatbot to talk about mental disorders with spanish teenagers. *International Journal of Human–Computer Interaction*, page 1–17.

National Institute of Mental Health. 2024a. Anxiety disorders. Web page, National Institute of Mental Health. Last reviewed December 2024; Accessed 16 June 2025.

National Institute of Mental Health. 2024b. Phobias and phobia-related disorders. Brochure / Fact Sheet, National Institute of Mental Health. Last reviewed December 2024; U.S. Department of Health and Human Services, National Institutes of Health; Accessed 16 June 2025.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy. Association for Computational Linguistics.

Shaina Raza, Deepak John Reji, Femi Shajan, and Syed Raza Bashir. 2022. Large-scale application of named entity recognition to biomedicine and epidemiology. *PLOS Digital Health*, 1(12):e0000152.

Reddit community r/emetophobia. 2025. r/emetophobia: Community for emetophobia support. Online forum (Reddit). Accessed 16 June 2025.

Reddit, Inc. 2025. Reddit api documentation. Web page, Reddit Developer Platform. Version 0.11; Accessed 17 June 2025.

Nils Reimers, Iryna Gurevych, and Kyunghyun Cho. 2022. all-minilm-l6-v2: Sentence transformer model. Hugging Face model card. 384-dim sentence embeddings; Apache 2.0 license; Accessed 17 June 2025.

scikit-learn developers. 2025. sklearn.svm.svc — support vector classification. API documentation, scikit-learn. Version 1.7.0; Accessed 17 June 2025.

Tianlin Zhang, Annika M. Schoene, Shaoxiong Ji, and Sophia Ananiadou. 2022. Natural language processing applied to mental illness detection: a narrative review. *npj Digital Medicine*, 5(1):46.