The Visibility of Depression in Social Media: Mapping Symptoms to Linguistic Features

Ștefana-Arina Tăbușcă[1] **Ana Sabina Uban**[2,3] **Liviu P. Dinu**[2,3]

[1]Interdisciplinary School of Doctoral Studies [2]Faculty of Mathematics and Computer Science
[3]Human Language Technologies Research Center, University of Bucharest, Romania
stefana.tabusca@s.unibuc.ro auban@fmi.unibuc.ro ldinufmi.unibuc.ro

Abstract

Digital phenotyping research assumes that depression symptoms are detectable in people's written discourse, yet there is room to explore which specific symptoms leave linguistic traces and which remain invisible. In this paper, using matched clinical and social media data from 169 Reddit users (eRisk 2021), we construct a clinical symptom network from BDI-II responses and a symptom-language bridge matrix mapping each of the 21 BDI-II symptoms to 15 curated LIWC-22 linguistic features. After FDR correction, 37 significant associations emerge, revealing a divide between cognitive-affective symptoms (sadness, worthlessness, suicidality) that leave clear linguistic traces through *mental health vocabulary*, *anxiety words*, and *first-person pronouns*, while others, like vegetative symptoms (sleep, appetite, irritability, libido) appear less visible as captured by the selected linguistic features. These findings suggest that there might be dimensions of depression that are missed by popular methods of text-based depression monitoring in social media data.

1 Introduction

Depression is increasingly understood not as a monolithic condition but as a system of mutually reinforcing symptoms (Borsboom 2017; Scheffer et al. 2024). Network psychometrics has formalized this perspective, revealing that symptoms like worthlessness and sadness occupy central positions in clinical depression networks (Bringmann et al. 2014; van Borkulo et al. 2015). In our work, we also adopt the network perspective, acknowledging the mentioned entailment of depression as an interdependent structure of symptoms. While this approach is well-documented in clinical surveys such as Bringmann et al. (2014), it remains unexplored in the area of raw natural language. Given computational approaches such as NLP (Natural Language Processing) frameworks are being seen more and more as significant tools for interdisciplinary research areas such as psychology and sociology, we also employ them in this study to explore the question: how do the semantic patterns of symptom expression found in digital text compare to the results of clinical assessments? While recent work in the area on NLP in mental health has moved toward more nuanced approaches, including emotion and cognitive features as well as some exploration of relationships between them (Zhang et al., 2023; Uban et al., 2021; Zhan et al., 2023), psychologically informed linguistic analysis (confirming that depressed individuals exhibit distinct part-of-speech patterns and elevated first-person pronoun usage (Rude et al., 2004; Bucur et al., 2021) or transfer learning using interpretable linguistic features in an attempt to connect mental health disorder manifestations across disorders (Uban et al., 2022; Zogan et al., 2022), the dominant paradigm remains focused on detection and classification rather than on investigating which specific symptoms are linguistically expressible (Bucur et al., 2025). As such, instead of attempting to extract symptom scores from text, we adopt a bridge framework that uses two validated instruments: BDI-II item scores serve as the clinical symptom measure, while LIWC-22 categories serve as the linguistic measure. This paper's primary contribution consists of the systematic mapping between them: a 21 x 15 symptom-language association matrix that identifies the "linguistic visibility" of each depression symptom in the context of the selected linguistic framework.

2 Data and Methodology

2.1 Dataset

We use the eRisk 2021 depression dataset (Parapar et al., 2021), which provides matched clinical and social media data for individuals who participated in a depression detection task. Data was filtered to include only posts in English; thus, we obtained

Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2026), pages 355–361

a sample consisting of 169 users who completed the BDI-II and authored a total of 72,430 English-language Reddit posts (median 282 posts per user).

2.2 Clinical Symptom Network

The BDI-II (Beck et al., 1996) is a 21-item self-report instrument measuring depressive symptom severity on a 0-3 ordinal scale. The justification behind constructing a network using the symptom information offered by the questionnaire stems from the goal of a deeper understanding of symptom interaction, as well as a facilitation ground for the later comparison with linguistic features.

Items Q16 (sleep changes) and Q18 (appetite changes), which include bidirectional sub-items, were recombined using the maximum of directional scores, consistent with standard BDI-II scoring and published network analyses (Bringmann et al., 2014). We estimated a Gaussian Graphical Model (GGM) using the EBICglasso algorithm (Foygel and Drton, 2010) with polychoric correlations for ordinal data and the recommended tuning parameter $\gamma = 0.5$. Network stability was assessed via nonparametric bootstrap (1,000 resamples) and case-dropping bootstrap for centrality stability (Epskamp et al., 2018). Expected influence (Robinaugh et al., 2016) was used as the centrality index, as it accounts for edge signs; this is relevant for networks where negative edges represent suppressive symptom relationships.

2.3 Sentence Embeddings Validation

To confirm that the Reddit text data contains depression-relevant signal before conducting symptom-level analyses, we computed sentence embeddings (BAAI/bge-large-en-v1.5)[1] of all posts against BDI-II item descriptions and aggregated a composite severity score per user as the 90th percentile of post-level cosine similarities, weighted by first-person pronoun density. Spearman rank correlations were calculated for this composite with the BDI-II total score, obtaining a value $\rho = 0.463$ (95% CI: [0.337, 0.575]). This supports the idea that natural language Reddit text carries meaningful depression-related information and motivates the subsequent symptom-language bridge analysis.

2.4 Linguistic Features

For the linguistic features, we utilized the LIWC-22 tool (Boyd et al., 2022), which is designed to aid in the extraction of meaningful psychological, emotional and cognitive textual features. We selected 15 LIWC-22 categories based on published associations with depression: emotional dimensions (*sadness*, *anger*, *anxiety*, *positive emotion*), cognitive-linguistic style (*first-person pronouns*, *negation*, *cognitive processing*, *authenticity*), and other clinically related content (*death*, *fatigue*, *mental health*, *illness*), interpersonal markers (*conflict*, *profanity*), and evaluative language (*lack/absence*). We note that the curated feature set is weighted toward affective, cognitive, and self-referential categories, which may favor detection of cognitive-affective symptoms. Future work will test whether contextual embedding features or symptom-specific lexicons alter the visibility gradient. LIWC features were aggregated from post-level to user-level using first-person pronoun density as a self-attribution weight, following evidence that first-person pronoun usage indexes self-referential discourse relevant to depression (Tølbøll, 2019). A sensitivity analysis confirmed that results were robust to this weighting choice (weighted vs. unweighted bridge matrix correlation: $\rho = 0.896$), heatmap comparison available in Figure 1. Still, the extensive effects of the weights remain an open question for future exploration.

2.5 Symptom-Language Bridge Matrix

For each of the 21 BDI-II symptoms and 15 LIWC features, we computed Spearman rank correlations between clinical item scores and user-level linguistic feature scores, yielding a 21×15 association matrix (315 tests). Benjamini-Hochberg FDR correction was applied at $\alpha = .05$ to control for multiple comparisons.

3 Results

3.1 Clinical Symptom Network

The BDI-II GGM contained 100 non-zero edges (density = 0.476). The correlation stability coefficient for expected influence was CS = 0.438, proving stable compared to the recommended threshold of 0.25 for sufficient interpretability (Epskamp et al., 2018). Worthlessness (Q14) exhibited the highest expected influence (1.321), followed by loss of interest (Q12, 1.150), energy loss (Q15, 1.105), concentration difficulty (Q19, 1.077), and fatigue (Q20, 1.067). Libido loss (Q21, 0.450) and appetite changes (Q18, 0.590) were the least central. The network is visible in Figure 2.

[1] https://huggingface.co/BAAI/bge-large-en-v1.5

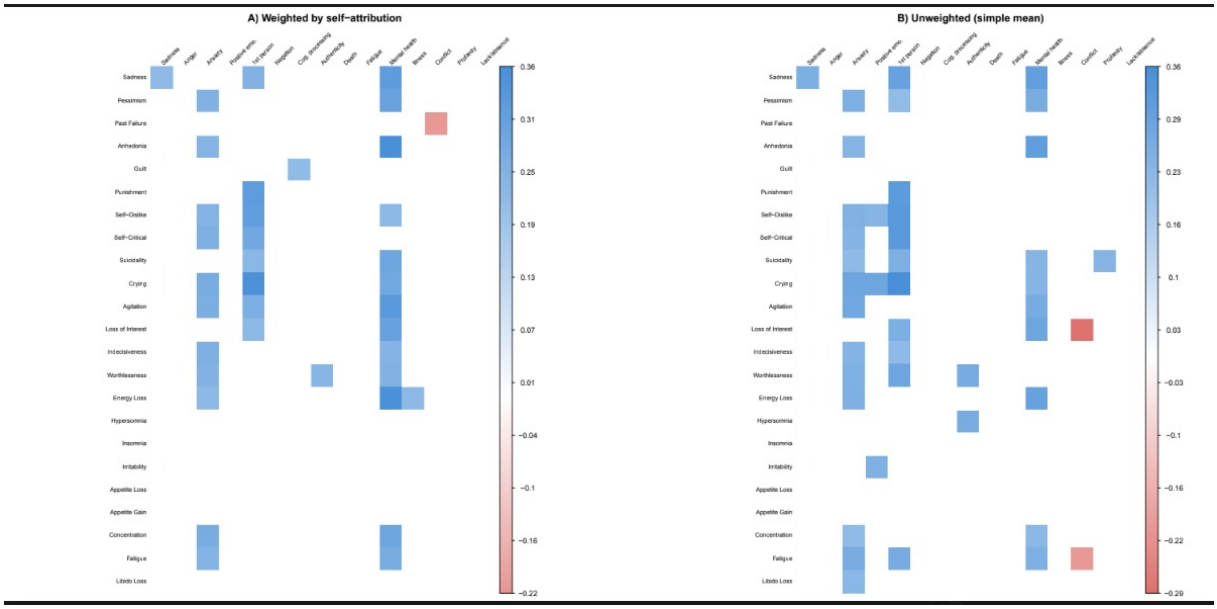

Figure 1: Comparison of correlations that passed FDR correction for weighted vs unweighted posts.

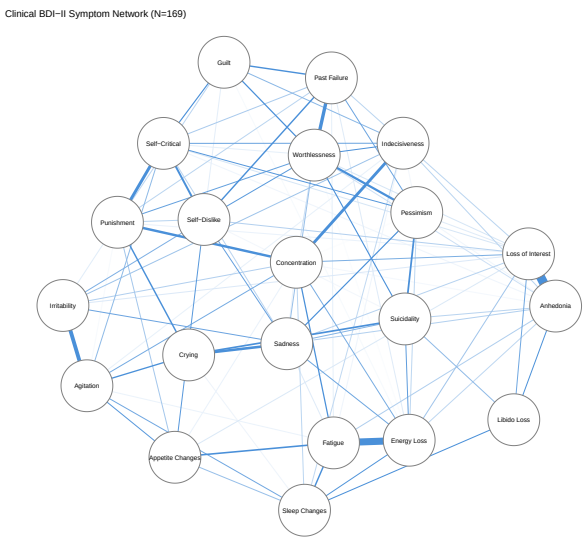

Figure 2: Clinical Symptom Network, with weighted edges representing the strength of the correlations.

The centrality of worthlessness is consistent with Beck's cognitive theory of depression (Beck et al., 1996), which positions negative self-evaluation as a core maintaining factor. The prominence of loss of interest and energy loss alongside cognitive symptoms suggests that in this sample, the clinical network is organized around cognitive-somatic symptoms, where internalized negative self-evaluation (worthlessness) connects to functional impairment (energy loss, concentration difficulty, loss of interest).

The finding that symptoms belonging to the vegetative group (libido loss, appetite changes) were the least central anticipates the linguistic visibility results that will be presented in the next subsection.

Comparing with published BDI-II networks, the centrality of worthlessness broadly replicates Bringmann et al. (2014), who found cognitive-affective symptoms forming the network's core. However, in our sample of Reddit users, energy loss and concentration difficulty show higher centrality than typically reported in clinical samples, potentially reflecting the functional demands of an online-active population where cognitive and energetic capacity are particularly prominent.

Community detection (Walktrap algorithm) revealed 5 communities, visible in Figure 3, which we can group as follows: 1 - Disregulated arousal (Irritability, Agitation, Concentration, Appetite Changes), 2 - Negative self-referential cognition (Worthlessness, Self-Critical, Guilt, Self-Dislike, Punishment, Past Failure, Indecisiveness), 3 - Affective despair (Sadness, Pessimism, Crying, Suicidality), 4 - Motivational withdrawal (Anhedonia, Loss of Interest, Libido Loss), 5 - Somatic-vegetative symptoms (Fatigue, Energy Loss, Sleep Changes).

The community structure further reveals that the linguistic visibility is not just a function of symptom severity or clinical centrality. The somatic-vegetative community, despite being peripheral in the clinical network (lowest expected influence scores for libido and appetite), is still associated with illness vocabulary, suggesting that somatic

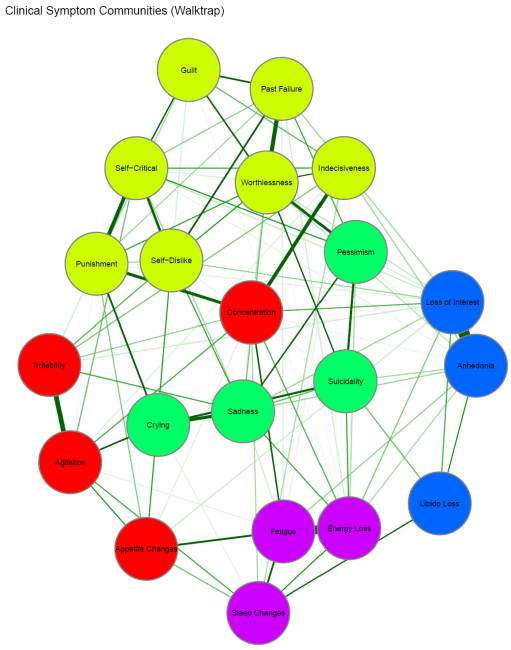

Figure 3: Clinical Symptom Network Communities computed through the Walktrap algorithm.

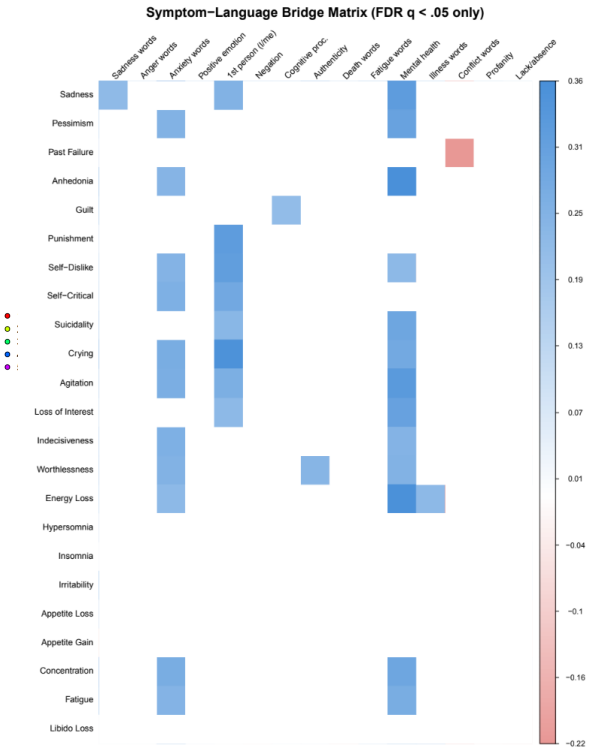

Figure 4: Symptom-language associations that survived FDR correction (q < .05)

experience surfaces in language through health-related rather than emotional or self-referential content. This might mean that targeted monitoring of illness and health language, rather than emotional sentiment alone, may be more informative for detecting somatic depression dimensions.

3.2 Symptom-Language Bridge Matrix

Of 315 symptom-language associations tested, 37 survived FDR correction (q < .05), involving 17 of 21 symptoms and 7 of 15 LIWC features, visible in Figure 4. The remaining four symptoms (sleep changes, irritability, appetite changes, and libido loss) had zero FDR-significant linguistic associations. All four belong to the group of vegetative or somatic symptoms.

Three LIWC features dominated the significant associations:

1. *Mental health vocabulary* (13 of 37 associations): Words like therapy, depressed, anxious were significantly associated with 13 symptoms, spanning both cognitive (pessimism, ρ = 0.302; worthlessness, ρ = 0.251) and somatic-adjacent (energy loss, ρ = 0.358; fatigue, ρ = 0.270) domains. This feature likely captures self-identified depression discourse rather than specific symptom content.

2. *Anxiety words* (11 of 37 associations): Anxiety language was significantly associated with 11 symptoms, most strongly with concentration difficulty (ρ = 0.269), crying (ρ = 0.268), agitation (ρ = 0.266), self-criticism (ρ = 0.258), and indecisiveness (ρ = 0.256). These form an anxious-ruminative cluster, consistent with the high comorbidity between depression and anxiety in clinical presentations (Hirschfeld, 2001).

3. *First-person pronouns* (8 of 37 associations): self-referential language was associated with some of the self-evaluative symptoms, such as crying (ρ = 0.356), punishment feelings (ρ = 0.324), self-dislike (ρ = 0.315), self-criticism (ρ = 0.284), agitation (ρ = 0.262) and suicidality (ρ = 0.234). The fact that first-person pronouns predict self-directed symptoms aligns with the literature on rumination, where excessive self-focused attention is a core aspect of depression (Spasojević and Alloy, 2001).

The only negative FDR-significant association was between *conflict words* and past failures (ρ = -0.219), suggesting that individuals reporting greater feelings of failure use fewer confrontational words; this is consistent with the idea of withdrawal rather than outward hostility. Additional significant singletons included *sadness words* with sadness

(ρ = 0.223), *authenticity* with worthlessness (ρ = 0.238), *illness words* with energy loss (ρ = 0.228), and *cognitive processing* with guilt (ρ = 0.216).

At the total-score level, the LIWC features most strongly associated with BDI-II total were *mental health vocabulary* (ρ = 0.404), *anxiety words* (ρ = 0.340), *first-person pronouns* (ρ = 0.285), and *illness words* (ρ = 0.210). Total post count was negatively associated with depression severity (ρ = -0.193, p = .012), consistent with behavioral withdrawal.

3.3 Symptom Community-Language Bridge Matrix

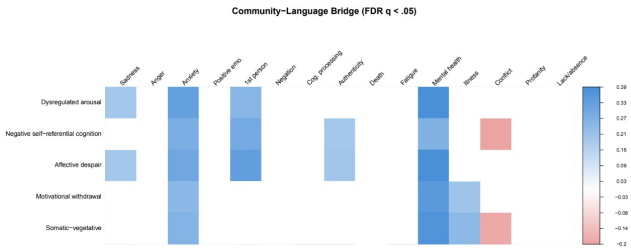

Figure 5: Symptom community-language associations that survived FDR correction (q < .05)

To examine whether symptom communities exhibit distinct linguistic profiles, we obtained composite BDI-II scores for each of the five Walktrap communities by averaging symptom scores for every symptom in a community. We then correlated them with LIWC features (5 x 15 matrix, FDR-corrected, visible in Figure 5). Of 75 associations, 21 survived correction. *Mental health vocabulary* and *anxiety words* were universally associated with all five communities, reflecting general depression discourse.

We also find several relevant associations at the community level which were not apparent at the individual symptom level: *first-person pronouns* and linguistic *authenticity* were associated with the affective despair (ρ = 0.330) and negative self-referential cognition (ρ = 0.293) communities, but not with somatic or motivational symptoms. *Illness vocabulary* is associated specifically with somatic-vegetative (ρ = 0.241) and motivational withdrawal (ρ = 0.205) communities but not cognitive ones. *Conflict language* was negatively associated with self-referential cognition (ρ = -0.203) and somatic-vegetative symptoms (ρ = -0.190), pointing once again to an opposition for external disputes. These values indicate that all symptom communities are linguistically detectable, but through different channels: cognitive-affective communities through self-referential language, and somatic communities through illness vocabulary.

4 Conclusions and Future Work

Our findings suggest that depression symptoms are not equally visible in Reddit online discourse in the explored setup of LIWC features. A pattern emerges from the matched clinical and linguistic eRisk data from the 169 Reddit users: cognitive-affective symptoms, especially those involving negative self-referential cognition, leave significant traces in natural language in social media, through mental health vocabulary, anxiety language, and first-person pronouns, while vegetative symptoms seem to be linguistically unexpressed. This asymmetry reflects a fundamental difference in how symptom types manifest in spontaneous discourse; the cognitive ones are mental states that people articulate, while the vegetative ones are bodily states that people experience.

A next step for this research would be to supplement LIWC features with contextualized representations, for example, using sentence embedding similarity between posts and symptom descriptions as a continuous feature or extracting symptom-specific linguistic markers.

The eRisk dataset consists of self-selected Reddit users posting in a narrative, anonymous environment. In this sense, Reddit discourse may systematically amplify cognitive-affective symptom expression relative to platforms with different affordances, such as structured forums or prompted clinical interviews. Replicating the bridge matrix across other platforms and datasets would clarify whether the observed findings reflect properties of depression symptom expression generally or of Reddit discourse specifically.

Ethical Considerations

All data were sourced from the eRisk 2021 dataset, which provides pre-anonymized social media posts. The dataset was obtained after signing a usage agreement. No attempts were made to deanonymize or contact individuals. The analyses presented are intended for research-level understanding of symptom-language associations and are not designed for individual clinical assessment or intervention.

Limitations

The eRisk sample consists of self-selected Reddit users, limiting generalizability to broader clinical populations. BDI-II scores reflect self-report, not clinical diagnosis. LIWC operates at the word level and cannot capture contextual meaning, negation, sarcasm, or indirect expression; symptoms expressed through complex narrative rather than individual words may be missed. The absence of significant LIWC associations for vegetative symptoms does not establish that these symptoms are inherently undetectable in text. Contextual embedding approaches, symptom-specific lexicons, or analysis of narrative structure and indirect expression could reveal linguistic markers of sleep disruption, appetite changes, or irritability that word-level features miss. Finally, the LIWC feature selection involved researcher judgment; a supplementary analysis using all available LIWC categories yielded consistent patterns but is not reported here due to space constraints. With 169 participants, confidence intervals for individual bridge correlations are wide, and replication in larger samples is needed.

Acknowledgements

This research was partially supported by the project "Romanian Hub for Artificial Intelligence - HRIA", Smart Growth, Digitization, and Financial Instruments Program, 2021-2027, MySMIS no. 334906

References

Aaron T Beck, Robert A Steer, Gregory K Brown, and 1 others. 1996. Beck depression inventory.

Denny Borsboom. 2017. A network theory of mental disorders. *World Psychiatry*, 16(1):5–13.

Ryan L Boyd, Ashwini Ashokkumar, Sarah Seraj, and James W Pennebaker. 2022. The development and psychometric properties of liwc-22. *Austin, TX: University of Texas at Austin*, 10.

Laura Bringmann, Lotte Lemmens, Marcus Huibers, Denny Borsboom, and Francis Tuerlinckx. 2014. Revealing the dynamic network structure of the beck depression inventory-ii. *Psychological Medicine*, pages 1–11.

Ana-Maria Bucur, Adrian Moldovan, Keerthi Parvatikar, Marcos Zampieri, Ashiqur R. KhudaBukhsh, and Liviu P. Dinu. 2025. On the State of NLP Approaches to Modeling Depression in Social Media: A Post-COVID-19 Outlook. *IEEE Journal of Biomedical and Health Informatics*, 29(6):4439–4451.

Ana-Maria Bucur, Ioana R. Podină, and Liviu P. Dinu. 2021. A Psychologically Informed Part-of-Speech Analysis of Depression in Social Media. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 199–207.

Sacha Epskamp, Denny Borsboom, and Eiko I. Fried. 2018. Estimating psychological networks and their accuracy: A tutorial paper. *Behavior Research Methods*, 50(1):195–212.

Rina Foygel and Mathias Drton. 2010. Extended bayesian information criteria for gaussian graphical models. *Preprint*, arXiv:1011.6640.

Robert M. A. Hirschfeld. 2001. The comorbidity of major depression and anxiety disorders: Recognition and management in primary care. *Primary Care Companion for CNS Disorders*, 3(6):24412.

Javier Parapar, Patricia Martín-Rodilla, David E Losada, and Fabio Crestani. 2021. Overview of erisk at clef 2021: Early risk prediction on the internet (extended overview). *CLEF (Working Notes)*, 1:864–887.

Donald Robinaugh, Alexander Millner, and Richard McNally. 2016. Identifying highly influential nodes in the complicated grief network. *Journal of Abnormal Psychology*, 125:747–757.

Stephanie Rude, Eva-Maria Gortner, and James W. Pennebaker. 2004. Language use of depressed and depression-vulnerable college students. *Cognition & Emotion*, 18(8):1121–1133.

Marten Scheffer, Claudi Bockting, Denny Borsboom, Roshan Cools, Clara Delecroix, Jessica Hartmann, Kenneth Kendler, Ingrid van de Leemput, Han Maas, Egbert Nes, Mark Mattson, Pat McGorry, and Barnaby Nelson. 2024. A dynamical systems view of psychiatric disorders—theory: A review. *JAMA Psychiatry*, 81:618–623.

Jelena Spasojević and Lauren Alloy. 2001. Rumination as a common mechanism relating depressive risk factors to depression. *Emotion (Washington, D.C.)*, 1:25–37.

Katrine Bønneland Tølbøll. 2019. Linguistic features in depression: a meta-analysis.

Ana-Sabina Uban, Berta Chulvi, and Paolo Rosso. 2021. An Emotion and Cognitive Based Analysis of Mental Health Disorders from Social Media Data. *Future Generation Computer Systems*, 124:480–494.

Ana-Sabina Uban, Berta Chulvi, and Paolo Rosso. 2022. Multi-Aspect Transfer Learning for Detecting Low Resource Mental Disorders on Social Media. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC 2022)*, pages 3202–3219.

Claudia van Borkulo, Lynn Boschloo, Denny Borsboom, B.W. Penninx, Lourens Waldorp, and Robert Schoevers. 2015. Association of symptom network structure with the course of depression. *JAMA Psychiatry*, 72.

Hongli Zhan, Desmond Ong, and Junyi Jessy Li. 2023. Evaluating subjective cognitive appraisals of emotions from large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14418–14446.

Tianlin Zhang, Kailai Yang, Shaoxiong Ji, and Sophia Ananiadou. 2023. Emotion fusion for mental illness detection from social media: A survey. *Information Fusion*, 92:231–246.

Hamad Zogan, Imran Razzak, Xianzhi Wang, Shoaib Jameel, and Guandong Xu. 2022. Explainable depression detection with multi-aspect features using a hybrid deep learning model on social media. *World Wide Web*, 25(1):281–304.