Automatic Annotation of Mental Health Recovery Narratives: A Benchmark Study

Shrankhla Pandey[1,2], Graham Murray[1], Ben Laws[1],
Stefan Rennick-Egglestone[3], Mike Slade[3,4] Sarah Morgan[1,2,5],
[1]Department of Psychiatry, University of Cambridge, UK.
[2]Department of Computer Science and Technology, University of Cambridge, UK.
[3]School of Health Sciences, University of Nottingham, UK.
[4]Health and Community Participation Division, Nord University, Norway.
[5]School of Biomedical Engineering and Imaging Sciences, King's College London, UK
Correspondence: sp2147@cam.ac.uk

Abstract

Manual annotation of mental health recovery narratives is slow and emotionally demanding, which limits the scalability of the digital mental health resource. A framework exists to characterise such narratives, called INCRESE, but there are currently no methods to automatically annotate the characteristics defined in INCRESE. We benchmarked the ability of support vector classifiers to annotate INCRESE characteristics when trained with three families of text representations: bag of words, GloVe static embeddings, and BERT contextual embeddings, using a dataset of 355 mental health recovery narratives. Characteristics related to diagnosis and turning points achieved a balanced accuracy greater than 0.67. Characteristics related to content warnings achieved a balanced accuracy of 0.72 but showed poor recall, which may be harmful for readers because it could lead to unsolicited exposure to sensitive content such as abuse or sexual violence. The lived-experience advisors endorsed the project objectives and addressed challenges of characteristic prioritization, adding insights not visible from quantitative metrics alone.

1 Introduction

Mental health recovery narratives can be defined as first-person non-fiction accounts of recovery from mental health issues (Ali et al., 2024); hereby referred to as recovery narratives. Slade et al. (2024) reported that access to online recovery narratives improved the quality of life for people affected by mental health problems. To make such narratives usable at scale, they must be systematically characterized. These characteristics help readers to search for recovery narratives that match their needs and circumstances (Llewellyn-Beardsley et al., 2025).

Conceptual frameworks are used to define such characteristics; one such framework is the Inventory of Characteristics of Recovery Stories (INCRESE) (Llewellyn-Beardsley et al., 2020). It characterizes narratives across multiple dimensions such as tone, trajectory, and content warnings; each INCRESE characteristic has one correct class.

In this paper, we develop the first benchmark for automatic annotation of INCRESE characteristics on a collection of mental health recovery narratives. We report the classification performance on 67 INCRESE characteristics and investigate misclassifications. This work is complemented by input from a group of lived-experience advisors who shared their opinions on the automatic annotation of recovery narratives.

2 Background

2.1 Recovery as a concept

The term 'recovery' is defined in this work as 'a way of living a satisfying life, with or without the continuation of mental health problems' (Research Into Recovery, b). It is grounded in narrator-defined progress and personal growth, rather than only clinical measures such as symptom scores or functional status. Slade et al. (2021) reported that recovery narratives can reinforce the effectiveness of existing clinical practices, by reducing communication barriers and normalizing mental health problems.

2.2 NEON Programme

Narrative Experiences Online (NEON) is a digital mental health intervention designed to deliver lived-experience narratives safely and at scale to the people currently experiencing mental health problems and their informal carers. To facilitate this, the NEON team curated, annotated, and utilized the NEON Collection, which includes recovery narratives sourced from individual donations and public and private collections.

A randomized controlled trial (Slade et al., 2024) of the NEON intervention for people with mental health problems found that access to online

Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2026), pages 100–118

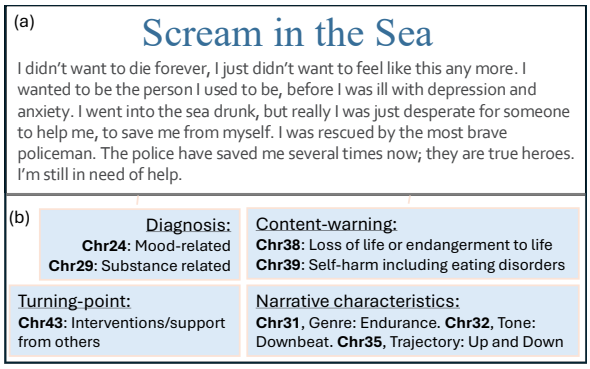

Figure 1: (a) An example of mental health recovery narrative from the NEON collection. (b) The blue boxes show the results of annotation with the INCRESE characteristics (Chrs) along with their associated higher-level section (underlined): Diagnosis, Content Warnings, Turning Point, and other Narrative characteristics. The story is a part of the NEON collection (Slade et al., 2021)

recovery narratives improved quality of life and increased meaning in life, among individuals with no recent (five-year) history of psychosis.

2.3 INCRESE tool

To make narratives searchable and safe, NEON relies on INCRESE, Inventory of Characteristics of Recovery Stories (Research Into Recovery, a), which defines 77 characteristics spanning narrative eligibility, mode, form, content warnings, turning points, and narrator attributes. Each characteristic has one correct class. INCRESE is intentionally descriptive rather than evaluative: it does not judge the quality or value of a narrative, but summarises how it is told and what it contains.

We focused on predicting 67 INCRESE characteristics from sections 3-7 of INCRESE. The first section, narrative eligibility, was excluded because we only had access to narratives which were eligible for inclusion. The second section, narrative mode, was excluded because it contained items such as whether the narrative contained text, audio or images and we only included text narratives under the scope of this work.

2.4 Annotation

Annotation assigns labels to data. Manual annotation uses human coders following predefined guidelines and is often used as a gold standard, but it is costly and time-consuming. Automatic annotation applies computational methods, such as machine learning, to label data at scale.

Manual annotation of recovery narratives: For the NEON intervention, recovery narratives were manually annotated using a 13-page guide on how to apply the INCRESE framework (Llewellyn-Beardsley et al., 2020). Only Section 5 (Content Warnings) was annotated by at least two coders; disagreements were discussed and resolved, though not documented. All other sections were annotated by a single coder, so inter-rater agreement is not applicable. Figure 1 (b) shows example annotations for the narrative 'Scream in the Sea'.

Need for automatic annotation: Manual annotation of recovery narratives is time-consuming and limits scaling up the collection. It requires coders to engage repeatedly and in depth with first-person accounts of adversity, struggle, and trauma. Such exposure patterns risk producing lasting psychological and emotional harm (Steiger et al., 2021). Longer narratives compound this burden, as annotation time scales with narrative length. In addition, the NEON intervention estimated the total cost of annotation at £21,840, a further breakdown of these costs can be found in Paterson et al.. These practical, psychological, and economic considerations underline the need for scalable, automated approaches to annotation.

3 Methods

Our objective was to systematically investigate the performance of automatic annotation for INCRESE characteristics, and to identify patterns in both the feasibility and failure of automatic annotation.

3.1 Dataset

The NEON collection contains over 600 narratives (NEON Collection). A subset of 357 narratives included consent from the narrators for wider research use and contained text, and this subset was used in our study. For many characteristics, the class distribution was highly imbalanced. This led to under-representation of some classes. We quantified class imbalance for each INCRESE characteristic using the imbalance ratio (He and Garcia, 2009), defined as:

$$\mathrm{IR} = \frac{n_{\text{largest class}}}{n_{\text{smallest class}}}$$

where $n_{\text{largest class}}$ is the number of samples in the largest class and $n_{\text{smallest class}}$ is the number of samples in the smallest class. Higher values of IR indicate greater class imbalance.

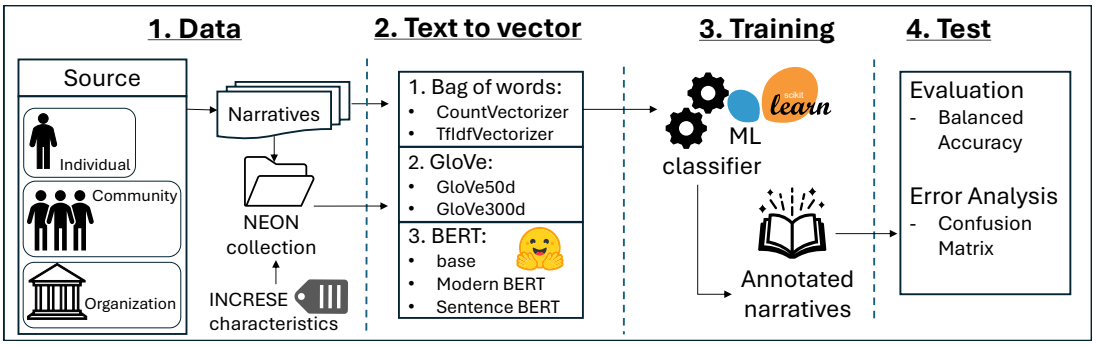

Figure 2: **Overview of our methodological pipeline.** (a). Narratives in the NEON collection were drawn from a range of individuals, communities and prior collections. The collection was manually labelled by the coders with INCRESE characteristics. (b). We converted text from the narratives to vectors, using a range of approaches. (c). The text vectors were input into a machine learning classifier to predict the classes of different INCRESE characteristics. (d). On test dataset, we report balanced accuracy. As part of error analysis we reported the confusion matrices.

3.2 Text preprocessing & representation

Some of the 357 narratives were mixed-media, containing images alongside text; in this study, we excluded all images and used only the textual content. After extracting the text, we applied preprocessing steps to remove artefacts including page numbers, special characters and emojis.

We removed two narratives whose lengths exceeded 3σ from the mean narrative length, where σ denotes the standard deviation of narrative length across the full collection. After exclusion, the remaining 355 narratives had a mean length of 1,202 words, with lengths spanning 86 to 3,387 words.

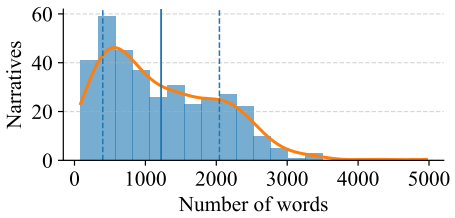

Figure 3: **Histogram of narrative lengths (word counts).** The solid line shows the mean, and dashed lines indicate ±1 standard deviation (σ). The y-axis represents the number of narratives.

To convert text into vectors, we used three types of approaches: bag of words methods, GloVe vector embeddings and BERT embeddings, see Table 1. These methods were chosen for the variety of information they embed and previous success in similar tasks (MacAvaney et al., 2021; Chan et al., 2025; Bucur et al., 2022; Bayram and Benhiba, 2022; Schmidt et al., 2025; Shreevastava and Foltz, 2021). In the first family of methods, bag-of-words, we used two methods: CountVectorizer, TfIdfVectorizer. In the GloVe family, we deployed two having different dimensions: GloVe50d and GloVe300d. In the BERT family, we used three methods: BERT-base, ModernBERT, and SentenceBERT. See Table 1 for a summary and the motivation for the use of these different approaches on our dataset. Using the text vectors extracted, we trained our machine learning classifiers.

3.3 Machine learning classification

Training setup: To benchmark, our focus is on feasibility; hence, we framed annotation for each INCRESE characteristic as an independent supervised text classification task. We first split the $n_{(sample)}$ = 355 narratives in a 3:1 ratio, with $n_{(train)}$ = 267 narratives in the training sample and $n_{(test)}$ = 88 narratives in the test sample. Due to heavy class imbalance, we applied inverse-frequency class weighting, upweighting minority classes proportional to $w_j = \frac{n_{(train)}}{k \cdot n_j}$, where k is the number of classes, and n_j is the number of samples in class j. We ran our experiments using logistic regression, random forest and support vector classifiers (SVC) using scikit-learn (Pedregosa et al., 2011) and observed comparable performance between the three approaches. We only report results from SVC here, for brevity.

For cross-validation, we used the narrative source as a grouping variable to enforce group-aware, stratified five-fold CV. Some sources contain inherent information about certain characteristics; for example, the Schizophrenia Oral History Project includes stories from people diagnosed with schizophrenia or schizoaffective disorder.

Representation	Description	Context window	Dim.	Justification for use in this work
Bag-of-words models (Pedregosa et al., 2011)				
CountVectorizer	Sparse token-count representation over a fixed vocabulary.	ℓ	$\lvert\mathcal{V}\rvert$	Baseline for the novel dataset, assessing lexical separability.
TF–IDF Vectorizer	Term-frequency inverse-document-frequency weighting to down-weight ubiquitous terms.	ℓ	$\lvert\mathcal{V}\rvert$	Reweighted baseline addressing vocabulary imbalance across narratives.
Static embeddings: GloVe (Pennington et al., 2014)				
GloVe (50d)	Pretrained distributional word embeddings derived from global co-occurrence statistics.	ℓ	50	Low-dimensional semantic and static representation; generalizes better than BoW to unseen text.
GloVe (300d)	Higher-dimensional variant capturing finer-grained lexical semantics.	ℓ	300	All benefits of GloVe50d, while testing whether extra representational capacity translates into measurable gains.
Contextual encoders (BERT family)				
BERT (base) (Devlin et al., 2019)	Bidirectional Transformer encoder producing contextualized token representations.	512 tokens	768	Contextual representation baseline for the annotation task.
ModernBERT (Warner et al., 2025)	Efficient long-context Transformer with larger window.	4096 tokens	4096	Long-context modeling of full-length narrative dependencies.
Sentence-BERT (MiniLM) (Reimers and Gurevych, 2019)	BERT fine-tuned for sentence-level embeddings.	512 tokens	384	Sentence-level representation for similarity-based annotation.

Table 1: Text representations explored in this work, their description, and motivations. ℓ denotes input text length; $\lvert\mathcal{V}\rvert$ is the vocabulary size.

We also performed hyperparameter optimization to avoid underfitting or overfitting of the SVC to optimize for balanced accuracy ; namely we tuned *C* [0.1, 1, 10, 100] to control regularization strength, the *kernel* [linear, rbf] to define linear or non-linear decision boundaries, and *gamma* [scale, auto, 0.1, 0.01] to adjust the smoothness of RBF (radial basis function) boundaries.

Evaluation: Given the severe class imbalance observed for many INCRESE characteristics, we evaluated the machine learning classifier performance using balanced accuracy (BA) on the test dataset. BA weights each class equally regardless of prevalence:

$$\mathrm{BA} = \frac{1}{C}\sum_{c=1}^{C}\frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FN}_c},$$

where C is the number of classes. For random chance, the expected BA is $1/C$, which provides a clear chance-level baseline.

3.4 Error analysis

Incorrect annotation carries a high risk of harm, especially in our dataset of recovery narratives. For readers, false positives can lead to invalidation. Someone looking for stories "like mine" may instead find narratives with incorrect labels and feel, "Even these people aren't like me.". For narrators, mislabelling can feel like being re-diagnosed without their consent. Missed warnings can cause distress, including dissociation or triggering reactions. For example, if Chr39 (self-harm, including eating disorders) is mislabelled, a reader who wants to avoid this content may still encounter harmful material (Yeo et al., 2021).

We analysed misclassifications using a confusion matrix. Every prediction a model makes is either correct or incorrect, and error can go in one of two directions. A false positive (FP) occurs when the model flags something as present when it is in fact absent, which can set incorrect expectations for the reader. A false negative (FN) occurs when the model misses a present instance and is unable to retrieve a relevant narrative. The false positive rate (FPR) captures how often absent instances are incorrectly flagged. The FPR is calculated by dividing the number of false positives by all cases that are truly absent. It represents how often narratives that lack a characteristic are incorrectly annotated

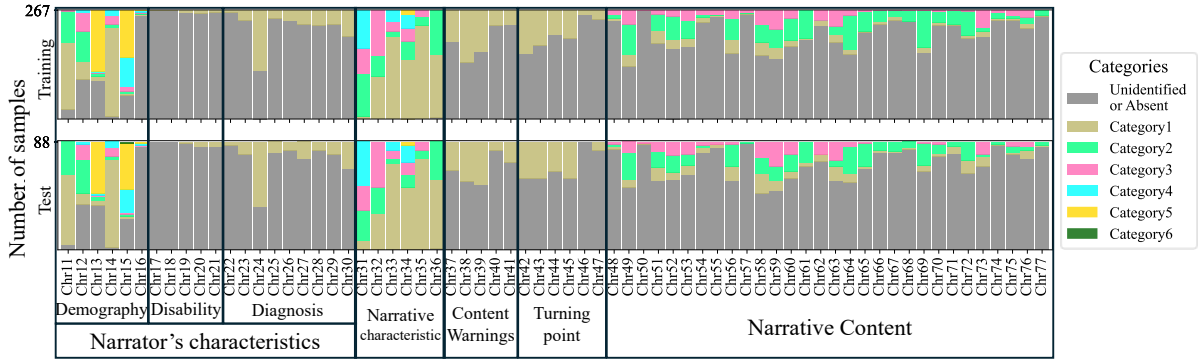

Figure 4: **Class imbalance across training and test datasets.** Grey bars represent samples in which the characteristic was absent or could not be identified from the narrative, illustrating the extent of class imbalance.

as having it. A high FPR implies that readers are presented with narratives that do not match the requested criteria. The false negative rate (FNR) captures how often present instances are missed. The FNR is calculated by dividing the number of false negatives by all cases that are truly present. It captures how often narratives that contain a characteristic are missed by the model. A high FNR indicates that relevant narratives are omitted from search results, reducing discoverability.

We selected the best performing model (highest BA) to investigate error patterns and report FP, FN, FPR and FNR.

4 Results and Discussion

4.1 Descriptive statistics

Training dataset: Some classes in the characteristics were severely under-represented in the training set (Figure 4), limiting the models' ability to learn stable decision boundaries for these classes. In particular, Chr13, Chr16, Chr17, Chr18, Chr50, Chr55, Chr57, Chr62, Chr66, Chr67, Chr68, Chr69, and Chr71 each contained fewer than three samples for at least one class. Under such extreme scarcity, supervised models fail to learn class representations that could be generalized (Banko and Brill, 2001). This effect is worse for binary-class characteristics, where missing or near-missing representation in one class effectively collapses the learning. We therefore did not train on binary-class characteristics Chr17 and Chr18 (narrator disability: visual and hearing difficulties, respectively).

Test dataset: The test set also contained characteristics that had no samples in at least one class. Specifically, Chr11, Chr16, Chr46, Chr55, Chr57, Chr62, Chr66, Chr67, Chr68, Chr71, Chr72, and Chr74 had zero samples in one of their classes. In these cases, balanced accuracy (BA) is mechanically deflated by 0.25 per missing class, irrespective of classifier behaviour. Chr46 ('Rude awakening') is a binary characteristic with one class missing, so its metrics reflect only a single class; we therefore excluded it from our analysis.

4.2 Machine learning performance

Figure 5 reports performances of SVC: one model per characteristic (64 characteristics) for each text representation method (7 methods), for a total of 448 models. We organise the characteristics by the number of classes. Within each class-count group, we sort characteristics by increasing imbalance ratio (IR), so the most balanced characteristic appears first and the most imbalanced appears last. For example, in the four-class group, Chr58 has (IR=4) while Chr50 has (IR=353).

A key takeaway from Figure 5 is that the seven text representations achieve broadly similar performance overall; no single representation dominates across characteristics. One likely reason is how we construct the BERT-based features. Although BERT models are reported to outperform BoW-based features (Bucur et al., 2022; Bayram and Benhiba, 2022; Schmidt et al., 2025), the BERT representations, by default, used the first (k) tokens (where (k) varies across BERT base, modern, and SentenceBERT; see Table 1). In addition, we used aggregated BERT vectors rather than stacked representations, which may reduce any advantage. A per-characteristic summary of which text representation performs best is provided in Figure 9, in appendix.

Chr19 (Mobility/stamina difficulties) and Chr21 (Self-care difficulties) illustrate the effect of severe class imbalance. Across all seven representations, models correctly predict the negative class but fail

Figure 5: **Machine learning performance for predicting narrative characteristics.** (a) Class imbalance ratios (green: 1–3, yellow: 3–10, red: >10). (b) Balanced accuracies for classifiers using different word representations; white indicates performance below the random chance, whereas progressively darker green shades indicate indicates stronger predictive performance.

to identify positive cases, yielding (BA=0.5). For Chr19, the test set contains only one positive instance; misclassifying this single example reduces balanced accuracy by 0.5, producing (BA=0.5) even when negative instances are handled correctly.

Setting difference between manual and automatic annotation: For manual annotation, coders used a 13-page guide (Recovery, 2025). In a few collections, narratives include photographs of the narrator. This visual context allows coders to directly infer demographic attributes, most notably age (Chr11), gender (Chr12), and ethnicity (Chr13), with relatively low ambiguity. As a result, human annotators operate with both multimodal evidence and task definitions whereas our automatic pipeline uses only text of the narratives. Therefore, discrepancies in performance, especially these characteristics do not solely reflect model error; they also arise from differences in the evidence available.

4.3 Error analysis

Our error analysis focused on two sections: narrator diagnosis and content warnings, as these were the sections where the models performed well enough to allow for a meaningful error analysis. In the remaining sections, model performance was inadequate to draw useful conclusions.

Narrator's Diagnosis: Figure 7 summarizes the confusion matrices for the eight diagnosis-related characteristics, Chr22 through Chr29. Here, we report metrics for each characteristic which had the highest BA. SentenceBERT performed best for 6 out of 8 characteristics, ModernBERT for Chr27, and BERT-base for Chr28.

We achieved specificity (TNR) greater than 0.82 for 7 out of 8 diagnostic characteristics. In these classes, false positives were comparatively rare, suggesting that the model typically avoids assigning a diagnosis when the corresponding condition is not present. Neuro-developmental, eating-related, mood-related, personality-related, obsessive–compulsive, psychosis-related, and trauma-related characteristics show good true-negative performance, reflecting a conservative decision boundary across most diagnoses. The Substance-related characteristic diverges from this pattern, with a substantially lower specificity (TNR = 0.64; TN = 50, FP = 28) but perfect sensitivity (TPR = 1.0; FN = 0).

Several characteristics, such as disorders related to personality, eating or psychosis, score low TPR despite high specificity, indicating that true conditions are sometimes under-detected. The mood-related characteristic, Chr24, is well balanced with IR close to 1 and has comparable sensitivity and

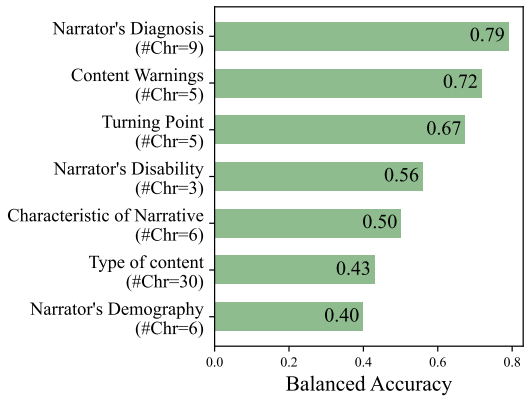

Figure 6: **Average performance for different sections**. #Chr shows the number of characteristics in the respective section.

specificity (both = 0.83), reflecting both its prevalence and the relative separability of its signals. Overall, for diagnosis-related characteristics of the narrator, the selected models minimize false annotations for most diagnoses, except for substance-related disorders.

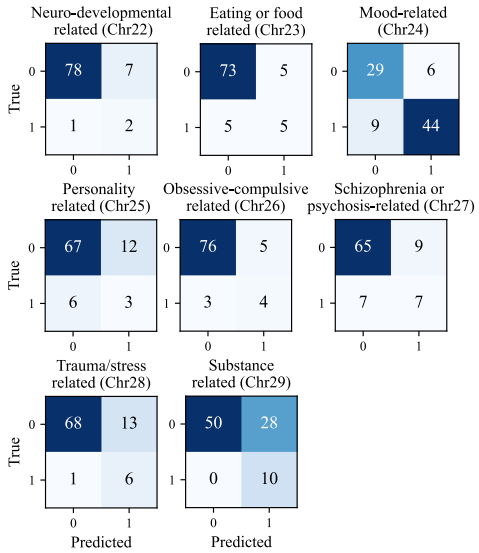

Figure 7: Confusion matrices for eight diagnostic-related characteristics (Ch22–Ch29).

Content Warnings in the narrative: Figure 8 presents the confusion matrices for the five content-warning characteristics, Chr37 through Chr41. For each characteristic, we report the model achieving the highest BA. Although the average BA for the content-warning section is 0.72 (Figure 6), this aggregate metric does not fully capture the severity of errors. As FN errors are particularly consequential for content warnings, we focus on the false negative rate (FNR). Only two out of five characteristics (Chr38 and Chr40) achieve an FNR below 0.3, indicating that, for several content-warning characteristics, a non-trivial proportion of relevant narratives are not flagged. The strongest performance is observed for Chr38, in where 25% of narratives where the narrator directly experiences or witnesses loss of life or endangerment to life are missed. The weakest performance occurs for Chr37, with an FNR of 0.48, meaning that nearly half of the narratives containing mentions of abuse or sexual violence are labelled as safe.

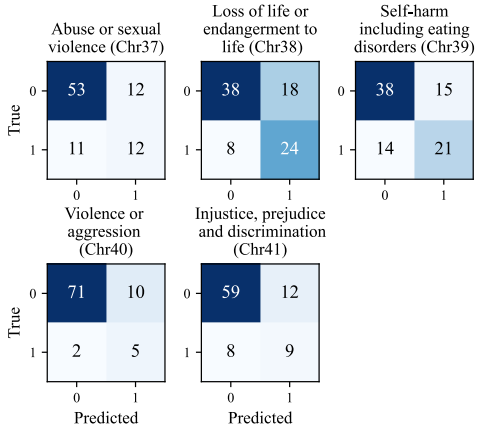

Figure 8: Confusion matrices for five content warning-related characteristics (Ch37–Ch41).

This behaviour reflects the decision to optimize for balanced accuracy during hyperparameter selection. Balanced accuracy gives equal importance to performance on each class, which provides a consistent basis for comparison across different INCRESE characteristics. However, this objective is not fully aligned with the priorities of the content-warning setting, where false negatives are more costly than false positives.

These results therefore establish a baseline under a general-purpose objective, while also indicating a clear direction for future work: re-optimizing these models using asymmetric loss functions or recall-oriented objectives to better reflect the safety-critical nature of content warnings.

4.4 Lived Experience Advisory Group

Quantitative performance alone does not determine whether automatic annotation is acceptable in practice. We therefore discussed our results with people with lived experience through DATAMIND's Lived Experience Advisory Group (LEAG) (DATAMIND, 2025a,b). We met with seven LEAG members, most of whom had prior experience reading recovery narratives online.

The members endorsed the broad premise and objective of scaling up the recovery narrative collection. They particularly welcomed the idea of using automated systems to detect content warning-related characteristics that appear consistently (72% of narratives contain one or more content warnings), such as mentions of violence, self-harm, or abuse. They further felt that such systems could reduce the emotional strain on coders by handling the initial pass, whilst ensuring that humans verify the final labels. The group did, however, express doubts about automatic annotation of subtler features such as tone, metaphor, identity shifts, and spiritual or existential experiences. Interestingly, our strongest performance (BA=0.86) was observed for spiritual or existential experiences (Chr45).

To understand why, we examined the Fisher ratios for Chr45 using the CountVectorizer model. The Fisher ratio (F) measures how well a word discriminates between classes by comparing the variance between class means to the variance within each class; a higher value indicates a more distinctive and reliable cue. The most discriminative words in the positive class include spiritual (F = 0.45), emergence (F = 0.19), divine (F = 0.17), spirit (F = 0.13), soul (F = 0.13), and consciousness (F = 0.13). These terms are lexically distinctive, providing strong surface-level cues that likely explain why this characteristic proved more tractable than the LEAG had anticipated.

When asked which INCRESE characteristics should be prioritised for automatic annotation, the LEAG advised against prioritisation, citing the likely high level of inter-individual variability in personal experiences of recovery. However, there was consensus that the level of accuracy would be variable. It would need to be strict for some characteristics (e.g. content warnings), whereas for other characteristics there's more room for error (e.g. tone).

This lived-experience panel endorsed the project objectives whilst surfacing challenges of prioritisation and adding insights not visible from quantitative metrics alone.

5 Related work

Much mental health NLP research focuses on detecting or predicting mental health conditions from text, primarily using social media data (Coppersmith et al., 2015; Zirikly et al., 2019; Tsakalidis et al., 2022; Tseriotou et al., 2025; Zanwar et al., 2023; Harrigian et al., 2021). These approaches frame the task as disorder detection and are typically designed to support screening or risk assessment. Recovery narratives differ fundamentally from this paradigm: narrators have already received a diagnosis, and rather than signalling the presence of disorder, their accounts describe long-term recovery processes across existing conditions including depression, anxiety, psychosis, schizophrenia, and personality and neurological disorders. They additionally offer multi-dimensional characterisation encompassing narrative structure, narrator perspective, content themes, and transformative moments. INCRESE was introduced to capture precisely these dimensions. Yet prior to our work, it had not been established how such rich, recovery-specific characteristics could be annotated or predicted at scale.

Previous research (Tian et al., 2024; Bae et al., 2025) has used computational methods for identifying story structure and turning points in fictional narratives; however, these methods have not been applied to mental health recovery narratives, where such elements are indicators of recovery pathways rather than merely narrative devices. Crucially, no prior work has adapted these methods to recovery narratives using a recovery-specific annotation framework.

Finally, although participatory and lived-experience involvement has been increasingly emphasized in mental health and AI research (Patrickson et al., 2023; Thompson et al., 2025), the resulting systems are rarely validated with input from those with lived experience.

This work filled these gaps by introducing an automatic annotation benchmark for recovery narratives grounded in INCRESE, systematically evaluating which recovery characteristics could be reliably predicted, and identifying practical limits to scaling recovery narrative collections.

Conclusion

We established baseline performance for automatic annotation of mental health recovery narratives using support vector classifiers. Our results show balanced accuracy above 0.8 for 6 INCRESE characteristics across diagnosis (4), content warnings (1), and turning points (1) but varies highly across sections and characteristics. Researchers considering automated characterisation should select optimisation metrics carefully, as different sections carry different misclassification consequences.

Limitations

The dataset size is relatively small for machine learning studies, which limits model learning and performance. While the dataset cannot be publicly released under its licence terms, we have provided dataset descriptions and a detailed evaluation protocol to mitigate reproducibility concerns.

In extracting text representations, we aggregate BERT family embeddings into fixed-length representations, which may dilute signal strength and potentially explain why these advanced models performed comparably to simpler models. This aggregation strategy may not optimally capture the

sequential and contextual information that transformer models are designed to encode.

The human coders had access to images as well as text, whereas ML algorithms only had access to text. Therefore the accuracies obtained by ML algorithms should be considered as a minimum achievable accuracy.

This work is the first benchmark study for automatic INCRESE annotation; further work is required to bring it closer to practical deployment in real-world settings for e.g. achieving higher recall for characteristics like content warnings, but precise thresholds have not been established. Development of large language models (LLMs) could be a fruitful avenue for further research. However, in our future work, we determine the ceiling performance that humans can achieve in annotating a subset of mental health recovery narratives from the NEON collection. More importantly we will examine which INCRESE characteristics are inherently subjective, with the aim of showing where automatic annotation could be trusted.

Ethics

Ethical approval for the curation of the Narrative Experiences Online (NEON) Collection was received from the London - West London and Gene Therapy Advisory Committee Research Ethics Committee in advance (18/LO/0991) on 9th July 2018. Specific consent for secondary data analysis has been documented for all NEON Collection narratives used in the current study. Nottinghamshire Healthcare NHS Foundation Trust were the sponsor for the NEON study. The sponsor approved the dataset licenses used to establish the terms of sharing NEON Collection narratives for the secondary analysis in the current study. Signed dataset licenses were in place before the transfer of narratives and narrative metadata for the purpose of secondary analysis.

Acknowledgments

The authors are grateful to the narrators who courageously shared their recovery journeys and granted permission for their stories to be used in this research, to the NEON team, and to the DATAMIND Lived-Experience Advisory Group. S.P. gratefully acknowledges PhD funding from the W.D. Armstrong Trust and both S.P. and S.M. received funding from the Accelerate Programme for Scientific Discovery, funded by Schmidt Futures. B.L. was supported by ImmunoMIND, UKRI. S.R.E. and M.S. were supported by the National Institute for Health and Care Research (NIHR) Nottingham Biomedical Research Centre (BRC) (NIHR203310). G.M. was supported by the NIHR Cambridge Biomedical Research Centre (NIHR203312) and the NIHR Applied Research Collaboration East of England.

References

Yasmin Ali, Stefan Rennick-Egglestone, Joy Llewellyn-Beardsley, Fiona Ng, Caroline Yeo, Donna Franklin, Elvira Perez Vallejos, Dror Ben-Zeev, Yasuhiro Kotera, and Mike Slade. 2024. Perception and appropriation of a web-based recovery narratives intervention: qualitative interview study. *Frontiers in Digital Health*, Volume 6 - 2024.

Suyoung Bae, Gunhee Cho, Yun-Gyung Cheong, and Boyang Li. 2025. CharMoral: A character morality dataset for morally dynamic character analysis in long-form narratives. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8809–8818, Abu Dhabi, UAE. Association for Computational Linguistics.

Michele Banko and Eric Brill. 2001. Scaling to very very large corpora for natural language disambiguation. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 26–33, Toulouse, France. Association for Computational Linguistics.

Ulya Bayram and Lamia Benhiba. 2022. Emotionally-informed models for detecting moments of change and suicide risk levels in longitudinal social media data. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 219–225, Seattle, USA. Association for Computational Linguistics.

Ana-Maria Bucur, Hyewon Jang, and Farhana Ferdousi Liza. 2022. Capturing changes in mood over time in longitudinal data using ensemble methodologies. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 205–212, Seattle, USA. Association for Computational Linguistics.

Callum Chan, Sunveer Khunkhun, Diana Inkpen, and Juan Antonio Lossio-Ventura. 2025. Prompt engineering for capturing dynamic mental health self states from social media posts. In *Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025)*, pages 256–267, Albuquerque, New Mexico. Association for Computational Linguistics.

Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. 2015. CLPsych 2015 shared task: Depression and PTSD

on Twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 31–39, Denver, Colorado. Association for Computational Linguistics.

DATAMIND. 2025a. Datamind: Better mental health through data-driven research. https://datamind.org.uk/. Accessed: 2026-01-22.

DATAMIND. 2025b. Lived experience advisory group. https://datamind.org.uk/patients-and-public/lived-experience-advisory-group/. Accessed: 2026-01-22.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Keith Harrigian, Carlos Aguirre, and Mark Dredze. 2021. On the state of social media data for mental health research. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 15–24, Online. Association for Computational Linguistics.

Haibo He and Edwardo A. Garcia. 2009. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284.

Joy Llewellyn-Beardsley, Yasmin Ali, Sylvia Bailey, Susie Booth, Paul Davis, Caroline Fox-Yeo, Donna Franklin, Julian Harrison, David King, Chris Newby, Fiona Ng, Scott Pomberth, Stefan Rennick-Egglestone, Dan Robotham, Roger Smith, Rianna Walcott, and Mike Slade. 2025. Lived experience narratives for mental health recovery: the narrative experiences online (neon) programme. Programme summary, Institute of Mental Health, University of Nottingham, Nottingham, UK.

Joy Llewellyn-Beardsley, Skye Barbic, Stefan Rennick-Egglestone, Fiona Ng, James Roe, Ada Hui, Donna Franklin, Emilia Deakin, Laurie Hare-Duke, and Mike Slade. 2020. INCRESE: Development of an inventory to characterize recorded mental health recovery narratives. *J. Recovery Ment. Health*, 3(2):25–44.

Sean MacAvaney, Anjali Mittu, Glen Coppersmith, Jeff Leintz, and Philip Resnik. 2021. Community-level research on suicidality prediction in a secure environment: Overview of the CLPsych 2021 shared task. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 70–80, Online. Association for Computational Linguistics.

NEON Collection. Neon collection. https://www.researchintorecovery.com/research/neon/neoncollection/.

Luke Paterson, Stefan Rennick-Egglestone, Sean P. Gavan, Mike Slade, Fiona Ng, Joy Llewellyn-Beardsley, Carmel Bond, Andrew Grundy, Joe Nicholson, Dania Quadri, Sylvia Bailey, and Rachel A. Elliott. 2022. Development and delivery cost of digital health technologies for mental health: Application to the narrative experiences online intervention. *Frontiers in Psychiatry*, Volume 13 - 2022.

Bronwin Patrickson, Mike Musker, Dan Thorpe, Yasmin van Kasteren, Niranjan Bidargaddi, The Consumer, and Carer Advisory Group (CCAG). 2023. In-depth co-design of mental health monitoring technologies by people with lived experience. *Future Internet*, 15(6).

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, and et al. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Research Into Recovery. 2025. Increse.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Research Into Recovery. a. Increse – inventory of characteristics of recovery stories: measure to characterise mental health recovery narratives. https://www.researchintorecovery.com/measures/increse/.

Research Into Recovery. b. Neon definitions: Key terms used in the narrative experiences online (neon) study. https://www.researchintorecovery.com/research/neon/definitions/.

Fabian Schmidt, Karin Hammerfald, Henrik Haaland Jahren, and Vladimir Vlassov. 2025. CFiCS: Graph-based classification of common factors and microcounseling skills. In *Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025)*, pages 106–115, Albuquerque, New Mexico. Association for Computational Linguistics.

Sagarika Shreevastava and Peter Foltz. 2021. Detecting cognitive distortions from patient-therapist interactions. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 151–158, Online. Association for Computational Linguistics.

Mike Slade, Stefan Rennick-Egglestone, Joy Llewellyn-Beardsley, Caroline Yeo, James Roe, Sylvia Bailey, Roger Andrew Smith, Susie Booth, Julian Harrison, Adaresh Bhogal, Patricia Penas Morán, Ada Hui, Dania Quadri, Clare Robinson, Melanie Smuk, Marianne Farkas, Larry Davidson, Lian van der Krieke, Emily Slade, and 7 others. 2021. Recorded mental health recovery narratives as a resource for people affected by mental health problems: Development of the narrative experiences online (neon) intervention. *JMIR Form Res*, 5(5):e24417.

Mike Slade, Stefan Rennick-Egglestone, Felix Ng, Jenny Yiend, Nicola Jones, Eleanor Longden, Victoria J. Bird, Mary Leamy, Jesper Larsen, Shaz Ali, Larry Davidson, Adam Bell, and Sarah Byford. 2024. Effectiveness and cost-effectiveness of providing recorded mental health recovery narratives to people with non-psychosis mental health problems (neon-o trial): A multicentre, parallel-group, superiority, randomised controlled trial. *The Lancet Regional Health – Europe*, 38:100805.

Miriah Steiger, Timir J. Bharucha, Sukrit Venkatagiri, Martin J. Riedl, and Matthew Lease. 2021. The psychological well-being of content moderators: The emotional labor of commercial moderation and avenues for improving support. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*, pages 1–14, New York, NY, USA. Association for Computing Machinery.

Alexandra Thompson, Victoria Bartle, Elizabeth A Remfry, Duncan J Reynolds, Michael R Barnes, Nick J Reynolds, and Barbara Hanratty. 2025. Public and patient involvement in artificial intelligence and big data healthcare research: An exploration of issues and challenges within the AI-Multiply project. *Health Expect.*, 28(6):e70490.

Yufei Tian, Tenghao Huang, Miri Liu, Derek Jiang, Alexander Spangher, Muhao Chen, Jonathan May, and Nanyun Peng. 2024. Are large language models capable of generating human-level narratives? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17659–17681, Miami, Florida, USA. Association for Computational Linguistics.

Adam Tsakalidis, Jenny Chim, Iman Munire Bilal, Ayah Zirikly, Dana Atzil-Slonim, Federico Nanni, Philip Resnik, Manas Gaur, Kaushik Roy, Becky Inkster, Jeff Leintz, and Maria Liakata. 2022. Overview of the CLPsych 2022 shared task: Capturing moments of change in longitudinal user posts. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 184–198, Seattle, USA. Association for Computational Linguistics.

Athanasia Tseriotou, Nikolaos Douzas, Molly Ireland, Ayah Zirikly, Dongji Yoo, Carlos Aguirre, Keith Harrigian, and Mark Dredze. 2025. Overview of the CLPsych 2025 shared task: Capturing mental health dynamics from social media timelines. In *Proceedings of the Eleventh Workshop on Computational Linguistics and Clinical Psychology*, pages 177–195, Abu Dhabi, UAE. Association for Computational Linguistics.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2526–2547.

Caroline Yeo, Stefan Rennick-Egglestone, Victoria Armstrong, Marit Borg, Donna Franklin, Trude Klevan, Joy Llewellyn-Beardsley, Christopher Newby, Fiona Ng, Naomi Thorpe, Jijian Voronka, and Mike Slade. 2021. Uses and misuses of recorded mental health lived experience narratives in healthcare and community settings: Systematic review. *Schizophrenia Bulletin*, 48(1):134–144.

Swamy Rakshith Zanwar, Daniel Wiechmann, Yuan Qiao, and Elma Kerz. 2023. SMHD-GER: A large-scale benchmark dataset for automatic mental health detection from social media in German. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1538–1557, Dubrovnik, Croatia. Association for Computational Linguistics.

Ayah Zirikly, Philip Resnik, Özlem Uzuner, and Kristy Hollingshead. 2019. CLPsych 2019 shared task: Predicting the degree of suicide risk in Reddit posts. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 24–33, Minneapolis, Minnesota. Association for Computational Linguistics.

A Best word representations for each characteristic

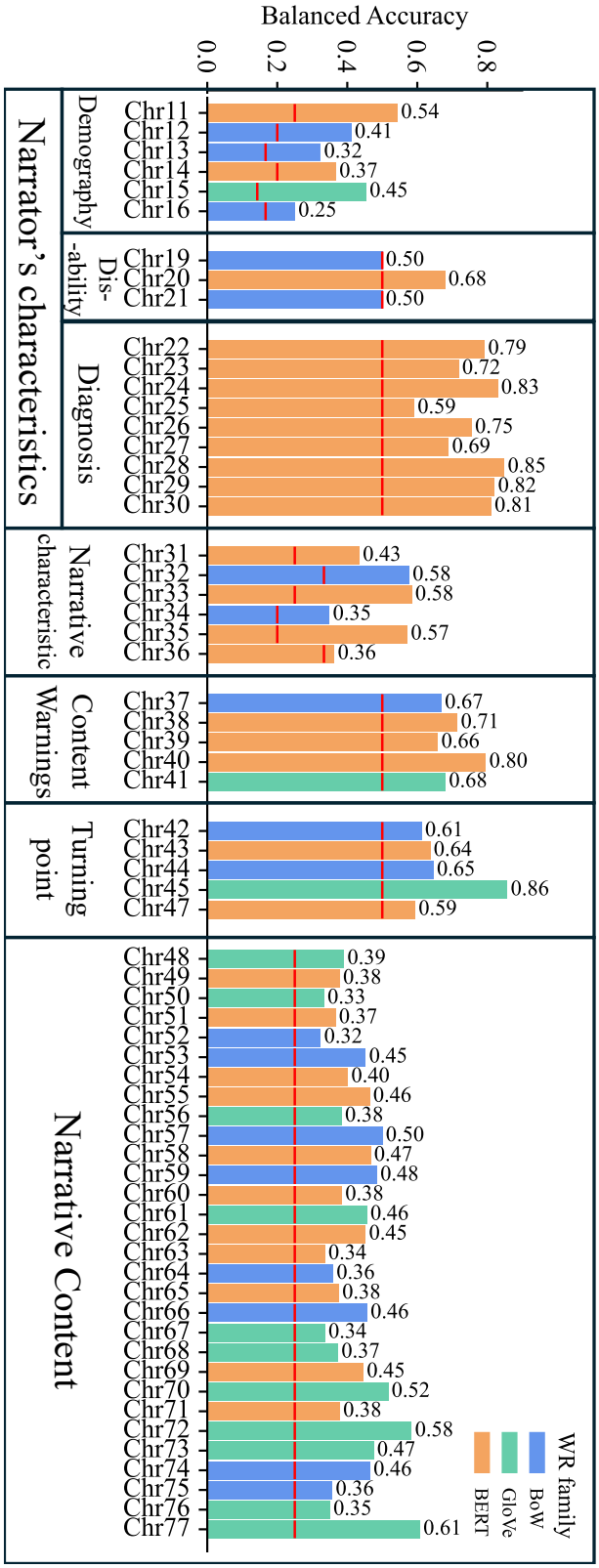

Figure 9: For each characteristic we show the model trained on the word representation having highest balanced accuracy.

B Description of all 67 INCRESE characteristics.

Table 2: Narrator's Demography characteristics.

Chr	Characteristic	Class
Chr11	Gender	0 = not identifiable 1 = female 2 = male 3 = other; Answer 'other' if the narrator identifies as e.g. trans, transgender, non-binary or any other gender identity.
Chr12	Age	0 = not identifiable 1 = 0–25 2 = 26–40 3 = 41–65 4 = 66+
Chr13	Ethnicity	0 = not identifiable 1 = Asian including Indian, Pakistani, Bangladeshi, Chinese 2 = African; Caribbean or any other Black, African or Caribbean background 3 = Dual/multiple ethnic group; Black, Caribbean and White; Black African and White; Asian and White, or any other dual or multiple ethnic background. 4 = Other ethnic group including Arab 5 = White including English, Welsh, Scottish, Northern Irish, British, Irish, Gypsy or Irish Traveller
Chr14	Stage of recovery	0 = Not identified; 1 = Working on recovery; Rebuilding/action/maintenance & growth stage: active engagement, rebuilding life, living beyond disability. 2 = Thinking about recovery; Awareness/contemplation/preparation stage: re-awakening of hope, awareness of dependency, making a decision to rebuild life. 3 = Not yet thinking about recovery; Moratorium/pre-contemplative stage: not currently engaged in thinking about or working towards recovery or wellbeing. 4 = Rejects recovery; Narrator explicitly rejects the concept of 'recovery' as used in mental health services.
Chr15	Location of narrator	0 = not identifiable 1 = Africa 2 = Asia 3 = Australasia 4 = Europe 5 = North America 6 = South America
Chr16	Sexuality	0 = not identified by the narrator 1 = Bisexual 2 = Gay man 3 = Heterosexual 4 = Lesbian/gay woman 5 = Other

Table 3: Narrator's Disability characteristics.
Each item is annotated as 0 = not identified, 1 = yes.

Chr	Characteristic	Description
Chr17	Visual difficulties	Blindness or partial sight. Do not tick solely based on narrator wearing glasses.
Chr18	Hearing difficulties	Deafness or partial hearing.
Chr19	Mobility / stamina difficulties	Difficulty e.g. walking, climbing stairs, lifting and carrying, or stamina or breathing difficulties.
Chr20	Cognitive difficulties	Difficulties with learning (e.g. intellectual disabilities), remembering (e.g. due to dementia), or concentrating.
Chr21	Self-care difficulties	Difficulty with self-care such as washing or dressing.

Table 4: Narrator's Diagnosis characteristics.
Each item is annotated as 0 = not identified, 1 = yes.

Chr	Characteristic	Description
Chr22	Neuro-developmental	e.g. Autism spectrum (ASD), Asperger's syndrome.
Chr23	Eating or food-related	e.g. anorexia, binge eating, bulimia.
Chr24	Mood-related	e.g. anxiety, bipolar 1 and 2, cyclothymia, depression, dysthymia, manic episodes, panic attacks, phobias, post-natal depression (PND).
Chr25	Personality-related	e.g. borderline personality disorder (BPD), narcissistic personality disorder, paranoid personality disorder.
Chr26	Obsessive-compulsive related	e.g. obsessive-compulsive disorder (OCD), body dysmorphic disorder.
Chr27	Schizophrenia or other psychosis-related	e.g. schizophrenia, delusional disorder, schizo-affective disorder, hearing voices.
Chr28	Trauma/stress-related	e.g. adjustment disorder, acute stress disorder, post-traumatic stress disorder (PTSD).
Chr29	Substance-related	e.g. problematic use of alcohol, cannabinoids, cocaine, other stimulants, opioids, hallucinogens, sedatives, solvents.
Chr30	Uses a non-diagnostic framework	Actively rejects use of diagnosis OR uses other terms instead of referring to diagnosis, such as hearing voices, trauma-based distress, being a survivor of services, self-identifies as Mad, OR uses an alternative explanatory framework, such as spiritual emergency. Do not tick just because items 22 to 29 are not ticked.

Table 5: Narrative characteristics.

Chr	Characteristic	Class
Chr31	Genre	1 = Escape; Main focus is on narrator's escape from or resistance to factors identified as preventing their recovery, e.g. difficult family or other circumstances, damaging or unhelpful services/treatment, forms of injustice, disputed diagnosis, rejection of diagnostic framework. 2 = Endurance; Main focus is on narrator's survival and ability to keep going despite prolonged conditions of e.g. distress, loss, trauma or difficult circumstances. Circumstances may be ongoing or in the past.

Chr	Characteristic	Class
		3 = Endeavour; Main focus is on narrator's ways of maintaining their recovery or achieving wellbeing, e.g. activities/attitudes/learning/relationships which help, ways of managing or restoring order, achievements or attaining goals. 4 = Enlightenment; Main focus is on narrator's transformation of their understanding or perspective of their situation, e.g. finding a new explanatory framework; their own empowerment or self-actualisation; a quest or journey of exploration; a sense of redemption through something greater than self, either humanistic or spiritual.
Chr32	Positioning	1 = Within services; Mental health treatments/services described on the whole as positive factors in narrator's recovery e.g. through treatments or services that worked, having positive relationships with mental health workers, or through delivering services (e.g. employment as a peer support worker). 2 = Despite services; Mental health treatments/services described on the whole as negative factors in narrator's recovery, actively preventing recovery or making mental health distress worse e.g. abusive treatment, negative staff attitudes, loss of rights or dignity. 3 = Outside the services; Mental health treatments/services are not mentioned, e.g. narrator recovered through other means, or accessed alternative treatments/services; or they feature only minimally e.g. narrator found treatments/services neutral or ineffective but not damaging.
Chr33	Tone	1 = Upbeat; Positive states e.g. buoyant, content, hopeful, proud, optimistic, reflective. 2 = Downbeat; Negative states e.g. agitated, apologetic, frenetic, pessimistic, sad, shaken, self-critical. 3 = Critical; Provocative or stimulating, e.g. angry, challenging, defiant, protesting. 4 = Neutral; Flat, e.g. matter of fact, disenfranchised, resigned, e.g. 'this is just how it is'.
Chr34	Relationship with recovery	1 = Living well; Recovery/sense of wellbeing is described as well-established, e.g. narrator is mainly experiencing wellness, is confident that future mental health problems will be manageable. 2 = Making progress; Recovery/sense of wellbeing is described as improving e.g. narrator is experiencing wellness and also regularly experiencing times of distress, is fearful that future mental health problems may not be manageable. 3 = Surviving day to day; Recovery/sense of wellbeing is described as challenging or tentative, e.g. narrator is focused on getting through a situation, is persistent in the face of very difficult circumstances, expects mental health problems to continue in the future. 4 = Recovered; Mental health problems are described mainly in the past tense, e.g. narrator does not expect to experience them in future; speaks of 'being recovered'. 5 = Rejects recovery; Narrator rejects the concept of 'recovery' as used in mental health services.
Chr35	Trajectory	1 = Upward; Shape of the narrative overall is an ascending progression towards recovery or wellbeing. Narrative may contain setbacks/periods of distress but focus is on progression.

Chr	Characteristic	Class
		2 = Up and down; Shape of the narrative overall is one of both upturns towards health/wellbeing and downturns towards distress. Ups and downs may be dramatic/roller-coaster changes or more drawn out over time but narrative is more or less equally distributed between ups and downs. 3 = Horizontal; Shape of the narrative overall contains no significant upwards or downturns, e.g. narrator may describe a sense of stagnating or taking one day at a time.
Chr36	Use of metaphor or symbolic language	0 = Unidentified 1 = Yes; If the narrative contains text or speech, is metaphorical or symbolic language used? Select yes for all image-based narratives. 2 = No

Table 6: Content Warning characteristics.
Each item is annotated as 0 = not identified, 1 = yes.

Chr	Characteristic	Description
Chr37	Abuse or sexual violence	Direct experiences or witnessing of any form of sexual, physical or emotional abuse, neglect, partner/domestic violence or acts of sexual violence, including e.g. rape or attempted rape, sexual assault, female genital mutilation (FGM), modern slavery, sex trafficking, child sexual exploitation, subjection to pornography or witnessing sexual acts, unlawful/inappropriate use of restraint, misuse of medication (e.g. over-sedation), forcible feeding or withholding food, enforced social isolation (preventing someone from e.g. accessing services or seeing friends), bullying, coercion, cyber-bullying, harassment, humiliation, intimidation, use of threats, verbal abuse.
Chr38	Loss of life or endangerment to life	Direct experiences or witnessing of e.g. admission to intensive care, bereavement, diagnosis of a life-threatening condition, loss of pregnancy, natural disaster, serious accident, suicide or attempted suicide, terrorist attack, torture, traumatic birth, traumatic termination of pregnancy, violent death of another, being threatened with a weapon, war/military combat.
Chr39	Self-harm including eating disorders	Direct experiences or witnessing of e.g. deliberate injury or harm to oneself, neglect of self, alcohol or substance misuse, eating disorder-related behaviours.
Chr40	Violence or aggression	Direct experiences or witnessing of e.g. acts of aggression, fights, rioting. .
Chr41	Injustice, prejudice and discrimination	Direct experiences, witnessing of or reference to e.g. hate speech, prejudice or discriminatory actions/behaviours/decisions on the basis of e.g. colour, disability, ethnic origin, gender identity, nationality, race, religion, sexual orientation. Experiences may be at individual/interpersonal level, organisational/institutional level (e.g. mental health services, prisons) or systematic/structural level (historical, cultural, legal, political or economic systems).

Table 7: Characteristics related to turning points in the narratives.
Each item is annotated as 0 = not identified, 1 = yes.

Chr	Characteristic	Description
Chr42	Taking charge	Change after taking charge e.g. of own illness, recovery process, problematic substance use, or own life generally. May be sudden and decisive or a longer process. May be accepting or rejecting help or treatment e.g. deciding to take or stop medication.
Chr43	Interventions/ support from others	Change after e.g. accessing helpful medication, treatment, groups or services (whether directly mental health-related or not), receiving support from family or friends, being confronted by family or friends.
Chr44	Self-acceptance	Change after e.g. an increase in confidence, growth in self-awareness, emotional release, moving away from internalised stigma. This may occur through own learning/inner work, involvement with support or other groups, or counselling/therapy.
Chr45	Spiritual/ existential experience	Change after e.g. finding a sense of meaning/purpose, experiencing a large shift in perspective, joining a spiritual/religious community, undertaking spiritual practices, conversion experiences, dreams, positive visions, being prayed for, support of a guide/teacher.
Chr46	'Rude awakening'	Change after a shock or realising how bad things have become, e.g. being admitted to hospital, being moved to a long-term ward, seeing negative effects on family members, the death of someone close, a suicide attempt.
Chr47	Shift in identity	Change after reclaiming a stigmatised identity e.g. based on ethnicity, gender, sexuality, mental health status. This may occur through finding others with similar experiences, own learning, own inner work or becoming involved in activism.

Table 8: Characteristics related to narrative content.
Each item is annotated as 0 = Not present, 1 = Present, 2 = Mainly positive, 3 = Mainly negative

Chr	Characteristic	Description
Chr48	Pregnancy/birth	Positive examples: wanted pregnancy, pregnancy as upward turning point Negative examples: traumatic pregnancies/birth experiences, unwanted pregnancy, abortion, post-natal difficulties, difficulties with getting pregnant or ability to have children.
Chr49	Family	Positive example: emotional or practical support or care from family members, relationships with family as source of happiness or resilience Negative examples: own or parents' divorce, separation, relationship breakdown or ongoing conflict, death or loss in the family.
Chr50	Being in care	Positive or negative experiences of adoption, living in foster care or non parental care.
Chr51	Education	Positive examples: having positive opportunities for learning/training, engaging in self-education/self-help. Negative examples: difficulties in school, school refusal, being excluded, not enjoying school being required to attend a course.
Chr52	Friendships	Positive examples: support of friends. Negative examples: breakdown of friendships, loneliness, social isolation, death of a loved one.
Chr53	Relationships	Positive examples: supportive partner, new relationship bringing hope. Negative examples: difficulties or conflict with partner.
Chr54	Housing	Positive examples: having own home, home being a source of pride or sanctuary. Negative examples: being vulnerably housed or homeless

Chr	Characteristic	Description
Chr55	Income	Positive examples: income enabling choices, improvement in income. Negative examples: debt, financial difficulties, poverty, experience of the benefits system.
Chr56	Work	Positive examples: providing meaning and purpose, developing positive identity. Negative examples: burnout, work-related stress, unstable employment, unemployment, unwanted retirement.
Chr57	Criminal Justice System	Positive examples: prison providing security, a turning point or opportunities. Negative examples: being arrested, being in prison or young offenders' institution, family member being in prison.
Chr58	Diagnosis	Positive examples: makes sense of a situation, is a relief, provides answers or sense of hope. Negative examples: disagreement with diagnosis, feeling stigmatised, loss of hope.
Chr59	Medication	Positive examples: felt better, relief, hope. Negative examples: felt worse, did not help, was not voluntary, had unwanted side effects.
Chr60	Relationship with mental health professional	Positive or negative experiences of relationships with e.g. community mental health team worker, crisis team, mental health nurse, psychiatrist, psychologist, social worker.
Chr61	Peer support	Positive or negative experiences of providing or receiving individual or group-based peer support. May be formal e.g. within mental health services or recovery colleges, or informal e.g. from support groups such as AA, Depression Alliance, Hearing Voices Network.
Chr62	Involuntary use of mental health services	Positive or negative experiences of experiences of e.g. being sectioned in hospital, compulsory medication, community treatment orders.
Chr63	Hospitalization	Positive or negative experiences of voluntary or involuntary psychiatric hospitalisation.
Chr64	psychological services	Positive or negative experiences of experiences receiving or delivering talking therapies or counselling e.g. arts or creative, CBT, DBT, group therapy, psychotherapy, person-centred, psychoanalytic, psychodynamic, solution-focused.
Chr65	Alternative therapies/healing	Positive or negative experiences of e.g. acupuncture, homeopathy, massage, meditation, Reiki, traditional Chinese medicine (TCM).
Chr66	Being in natural environments	Positive or negative experiences of the natural environment or outdoor activities, green therapy.
Chr67	Animals or pets	Positive or negative experiences of contact with animals including with therapy / emotional support dog, or care of a pet.
Chr68	Community activities	Positive or negative experiences of engagement with community-based groups or organisations, whether or not mental health-related.
Chr69	Hobbies, interests, creative activities	Positive or negative experiences of leisure activities, hobbies, recreational interests or creative activities e.g. art, crafts, music (listening or playing), performance, reading, dance, writing.
Chr70	Physical activities	Positive examples: gaining or maintaining fitness, being involved in sports or other physical activities. Negative examples: being unable to participate in such activities, participation leading to feeling worse.
Chr71	Activism	Positive examples: participation in political or other activities to achieve change in a group, organisation, service or system. Negative examples: participation in the above leading to e.g. burnout, being overwhelmed.

Chr	Characteristic	Description
Chr72	Spiritual or religious activities	Positive examples: attending a religious/spiritual community, group or place of worship, articipating in spiritual practices e.g. prayer, meditation, retreats. Negative examples: above activities as compulsory or contributing to mental health distress.
Chr73	Stigma	Positive examples: rejecting or overcoming stigma or shame due to mental health issues or for other reasons, e.g. ethnicity, gender. Negative examples: experiencing disapproval of others, or own shame due to mental health issues or other reasons.
Chr74	Caring responsibilities	Positive example: care of someone else giving a sense of pride. Negative example: care of someone else worsening own mental health problems.
Chr75	Family experiences of mental health issues	Positive example: mental health issues of other family members leads to a shared sense of understanding. Negative example: mental health issues of other family members has detrimental effect on narrator's own mental health.
Chr76	Diet or nutrition	Positive or negative experiences of diet or nutrition linked to improving, maintaining or worsening narrator's mental health.
Chr77	Volunteering	Positive or negative experiences of any unpaid work.