Differential Diagnosis of Borderline Personality Disorder: Machine Learning Models of Subjective Experiences

Adam Horvath, Mark Dras, Miriam K. Forbes

Macquarie University

Author Note

Adam Horvath, Department of Psychology, Macquarie University; Mark Dras, Department of Computing, Macquarie University; Miriam K. Forbes, Centre for Emotional Health, Department of Psychology, Macquarie University.

Adam Horvath is now at the Clinical Psychology Unit, University of Sydney.

Correspondence concerning this article should be addressed to Adam Horvath, Clinical Psychology Unit, University of Sydney, Camperdown, 2050, New South Wales. E-mail: adam.horvath@sydney.edu.au

Abstract

The present study aimed to capture the subjective experiences that distinguish BPD from closely related mental disorders to aid differential diagnosis. Posts from Reddit were downloaded from seven mental health discussion groups. The topics discussed in each post were extracted using a combination of machine learning approaches. Logistic regression models were used to classify whether posts originated from the BPD support group versus other support groups. The average classification performance was well above chance, even relying on only 25 subjective experiences, and after excluding topics that were objective markers of diagnoses (e.g., names of the disorders or medications). Fear of abandonment emerged as the key differentiator of BPD, and 11 of the 25 topics related directly to the definition of BPD in the DSM-5. However, several of these were typical of other disorders as well, raising questions about their diagnostic utility.

Differential Diagnosis of Borderline Personality Disorder: Machine Learning Models of Subjective Experiences

Accurate and timely diagnoses for mental disorders are important for connecting individuals with effective treatment. However, people with Borderline Personality Disorder (BPD) have to wait 15 years on average before receiving a diagnosis (Ng, Townsend, Miller, Jewell, & Grenyer, 2019). Moreover, practitioners often report uncertainty about the accuracy of their diagnosis of BPD (Sisti, Segal, Siegel, Johnson, & Gunderson, 2016) despite the detailed diagnostic criteria available in traditional classification systems, such as the Diagnostic and Statistical Manual of Mental Disorders (DSM-5; American Psychiatric Association, 2013). This study aims to improve the diagnosis of BPD by characterising the most distinctive subjective experiences of BPD, compared to other closely related disorders, which may aid in differential diagnosis.

(Mis)Diagnosing Borderline Personality Disorder using the DSM-5

BPD is the most prevalent personality disorder in the DSM-5 (1-4% community prevalence rate; Australian Bureau of Statistics, 2019; Leichsenring, Leibing, Kruse, New, & Leweke, 2011; Paris, 2010). The DSM-5 diagnostic criteria include a fear of abandonment, unstable relationships, identity disturbance, impulsivity, suicidal and self-harming behaviour, mood instability, feelings of emptiness, anger, and dissociative symptoms (American Psychiatric Association, 2013). Of these nine criteria, self-harm, and suicidality are of special concern: people with BPD are 50 times more likely to die by suicide than people in the general population (Cristea et al., 2017). Furthermore, due to recurrent self-harming and suicidal behaviours, people with BPD often require emergency department visits and hospitalisation (Leontieva & Gregory, 2013; Warrender, 2015). Although BPD is treatable with specialised

psychotherapy (Cristea et al., 2017), the long delay before receiving a diagnosis leads to ongoing distress and impairment among sufferers, burdens on the healthcare system, and potential lives lost to suicide, highlighting the need for more timely diagnoses (Ng et al., 2019). A correct and timely diagnosis is also critical both to validate sufferers' experiences so they can start a recovery journey (Ng et al., 2019), and to facilitate the specialised treatment that BPD requires, such as dialectical behaviour therapy or psychodynamic approaches (Campbell, Clarke, Massey, & Lakeman, 2020; Cristea et al., 2017).

Further to the issues with delayed diagnosis, misdiagnosis presents another concern for people with BPD, given its substantial symptom-level and conceptual overlap with at least six other disorders, as explored further below: generalised anxiety disorder, antisocial personality disorder, bipolar disorders, major depressive disorder, narcissistic personality disorder, and schizophrenia (Beatson, Broadbear, Duncan, Bourton, & Rao, 2019; Kotov et al., 2017; Ng et al., 2019; Renaud, Corbalan, & Beaulieu, 2012; Stern & Yeomans, 2018). One of the key challenges leading to misdiagnosis is the difficulty in delineating BPD from these disorders (Bayes, Parker, & Paris, 2019). For instance, according to the DSM-5, both BPD and antisocial personality disorder are characterised by manipulative behaviour (American Psychiatric Association, 2013). Similarly, both BPD and bipolar disorders are cyclothymic (Bayes et al., 2019).

More broadly, the DSM-5 relies on the assumption that all nine criteria of BPD are equally descriptive of the disorder (i.e., unranked), and that meeting any five or more would yield the same diagnosis. However, this is unlikely to be true: for instance, Linehan (1993) argued that anger, a diagnostic criterion of BPD since the DSM-III (American Psychiatric Association, 1980), stems from fear of abandonment and desperation, which raises the question of whether anger is as independently evaluable and equally descriptive of BPD as the

other eight criteria. Furthermore, some criteria are more important than others for the differential diagnosis of BPD. For example, Bayes et al. (2016) showed that the key differentiators between BPD and bipolar disorders were relationship difficulties and sensitivity to criticism, while other diagnostic criteria from the DSM-5 did not emerge as important predictors. These examples highlight the challenges of a consensus-based, rather than data-driven, definition of diagnostic criteria and have motivated researchers to look into newer, data-driven models of mental disorders.

Other approaches to defining BPD

In dimensional models, such as the Alternative Model for Personality Disorders in the DSM-5 (p. 761) and the Hierarchical Taxonomy of Psychopathology (HiTOP; Kotov et al., 2017), disorders are understood as a combination of traits (or dimensions) — similar to how personality is defined as the combination of five distinct factors in the Five-Factor Model (McCrae & Costa, 2003). Both the overlap and the differences between the disorders are captured in these dimensional models: for instance, in the HiTOP model, BPD shares the *distress* component with both major depressive disorder and generalised anxiety disorder, and *antagonism* with both antisocial and narcissistic personality disorders, acknowledging the similarities between these disorders.

The challenge with describing mental disorders with a limited number of broad dimensions is that finer, clinically relevant details can be lost. Livesley (2020) highlighted that the specific impairments that are the focus of treatment are best captured in the language of the clients themselves, rather than in the clinical terms defined in research: clinical terms, such as *neuroticism*, often do not map onto any phrases in everyday language. Subjective descriptions of psychopathology, however, are not typically considered when devising methods for diagnosing mental disorders. For example, structured clinical interviews and

diagnostic questionnaires are often direct translations of the diagnostic manuals, created and phrased by and for researchers and clinicians (e.g., DeShong, Mullins-Sweatt, Miller, Widiger, & Lynam, 2016; Hyler et al., 1988; Leichsenring, 1999). This often results in questions with a lot of "clinical phrases and jargon" (p. 175) and this language can create a distance between clinicians and patients that impedes the development of a strong therapeutic relationship (Beattie, Murphy, Burke, O'Connor, & Jamieson, 2019).

Given the challenges with using diagnostic models and the research-focused language of diagnostic questionnaires, clinicians often forego these formal definitions and instead use an intuitive (*prototypical*) approach when diagnosing people with BPD (Bayes et al., 2016; Livesley, 2020). This prototypical approach has advantages compared to counting how many independent diagnostic criteria a person meets in the DSM-5: personality disorders are pervasive by definition, rather than tied to distinct behaviours, times, or places, so a coherent pattern should be observable (American Psychiatric Association, 2013). On the other hand, this approach may be influenced by the biases of the clinician (Lilienfeld & Lynn, 2014).

In sum, identifying the distinguishing features of individuals' subjective experiences of BPD — in their own language — could represent a valuable resource for facilitating more accurate diagnoses. Machine learning represents an ideal framework in which to achieve this aim.

Machine Learning in Diagnostics

Machine learning is a data modelling technique that aims to create predictive models that can generalise beyond the sample in which they were created, without making distributional assumptions on the whole population (Ankam, 2016). These models generally include a large number of predictors and require large sample sizes; therefore, instead of statistical significance — which depends on the sample size — other metrics, such as the Area

Under a Receiver Operating Characteristic Curve (AUC), are reported (the sample size fallacy; Lantz, 2013).

Despite its increasing popularity, machine learning received two main initial criticisms in psychology research, mostly due to the differences between this modelling technique relative to the conventions of traditional inferential statistics (Siddaway, Quinlivan, Kapur, O'Connor, & de Beurs, 2020). First, machine learning often needs much larger sample sizes than inferential statistics (Ankam, 2016), which can be prohibitively difficult to achieve for some research questions (e.g., those focused on rare populations). However, this limitation is not relevant if data are available in large quantities (e.g., organically created datasets of online forum posts). Machine learning has been successfully applied to process these kinds of data in clinical psychology research, for instance, by predicting the diagnosis of a patient, or who would respond well to specific treatment options (Shatte, Hutchinson, & Teague, 2019). Second, complex models with a large number of predictors may be difficult to decipher. If the resulting models are too complex to interpret, researchers cannot verify how the input variables contributed to the calculated results (i.e., the model operated as a *black-box*; Siddaway et al., 2020). Given the drawbacks of black-box modelling, there is ongoing research on how to make black-box models more interpretable (Molnar, 2020). Regardless, Yin, Sulieman, and Malin (2019) recommended using machine learning models that yield interpretable results in psychology research, such as logistic regression.

Diagnosis and User-Generated Content

Personality disorders can be diagnosed using structured clinical interviews and self-report questionnaires (Samuel et al., 2013). Therefore, people's own description of their symptoms carries valuable information for diagnosis that can have direct relevance to the information clinicians have to work with (i.e., when talking with their patients). However, as

discussed above, there are differences in the language used by clinicians and patients, which means that a new modelling approach is needed to capture the experiences of people with BPD in their own words. Content created organically by people with BPD (*user-generated content*) describing their subjective experiences is thus a strong candidate data source for machine learning, as it comes directly from people's own description of their symptoms and can be readily found in vast quantities in online discussion forums.

To circumvent the limitations of other social media platforms, such as low signal-to-noise ratio (e.g., Facebook, Twitter; Ernala et al., 2019) researchers have turned to Reddit for its better-quality mental health-related content. Reddit is an online, anonymous discussion site, where users can self-organise into topic-specific discussions, called *subreddits*. Each subreddit is focused on a single topic: some are about specific cars or ancient art; others act as support groups for individuals' experiences related to particular mental disorders, such as BPD. Users on Reddit can vote each post "up" or "down" to rank its position compared to other posts (Choudhury & De, 2014). Importantly, users tend to upvote posts that matter to them and are relevant to their own subjective experiences (Kassaeyan, 2016, p. 47). Furthermore, the absence of upvotes has been found to signal off-topic posts (Guimaraes, Balalau, Terolli, & Weikum, 2019) — or, in the case of the present research, posts that were unlikely to reflect experiences with the disorder. The anonymity of Reddit also allows its users to talk more freely about their problems compared to Facebook users (Choudhury & De, 2014). Thorstad and Wolff (2019) showed that posts in mental disorder subreddits are almost exclusively about the individuals' subjective experiences with the disorders, and other studies have found that the topics of discussion in Reddit mental disorder support groups reflect clinically relevant content (e.g., Chakravorti, Law, Gemmell, & Raicu, 2018; Grant, Kucher, Leon, Gemmell, & Raicu, 2017; Thorstad & Wolff, 2019).

In sum, posts from Reddit represent an ideal data source for machine learning models aiming to characterise the distinguishing features between a set of related mental disorders. Machine learning research on user-generated content has shown promising results (e.g., Chang, Boyd-Graber, Gerrish, Wang, & Blei, 2009), but this type of research on Reddit support groups has not yet contributed to clinical psychology research. There are four particular methodological limitations that are likely to have introduced and compounded problems across several steps of data processing in such studies to date. First, all posts being treated as equal, even though some posts carried more relevant information about the disorders (e.g., ongoing distress about breaking up with a partner) while some did not (e.g., a photo and description of a cat). Second, relying too heavily on researchers' subjective interpretation of the machine learning results, which is vulnerable to bias and over-interpretation. Third, using inflexible semantic models, which do not cater for the colloquial, casual nature of user-generated content and are likely to lose much of the information contained in slang words and typographical errors that are typical to user-generated content. Finally, using black-box models that are often uninterpretable because they obscure how their answers are calculated and undermine the potential for these models to provide generalisable information that advances our understanding of mental disorders.

The Present Study

The present research modelled the differences between BPD, generalised anxiety disorder, antisocial personality disorder, bipolar disorders, major depressive disorder, narcissistic personality disorder, and schizophrenia using Reddit support group posts to answer the exploratory research question: What subjective experiences specifically differentiate BPD?

To overcome the four main methodological limitations of previous research, the present study incorporated a novel combination of machine learning techniques. First, posts with

relevant content about the disorders were identified and prioritised by taking upvotes into account. Second, highly homogenous keywords describing the same subjective experience (*topics*) were identified and named objectively based on prototypical keywords. This step was targeted at addressing the over-reliance on the subjective interpretation of researchers in past research, and to identify themes that characterised the content in the posts across the seven support groups. Third, a language modelling approach flexible enough to capture the semantics of everyday language was used. This was achieved by using a semantic-learning framework called FastText (Bojanowski, Grave, Joulin, & Mikolov, 2017), described in more detail below. Finally, this study used logistic regression models that maximised the interpretability of the results to predict which support group each post came from, based on the content in the post. Together, these approaches were used to create a ranked list of subjective experiences to differentiate between BPD and other disorders to show how typical or atypical each experience was to BPD.

It was hypothesised that the models would be able to differentiate between the support groups based on the subjective experiences described in each post. Specifically, the logistic regression models would be able to identify whether a post belonged to the BPD support group or another support group, as measured by a better-than-chance classification accuracy (i.e., AUC > .5).

Based on previous research, some topics were expected to reflect diagnoses or treatments of the disorder, rather than subjective experiences of BPD (*noise topics*). To clarify the most typical experiences of BPD, the models were also tested excluding the noise topics (i.e., removing topics that referred to the names of mental disorders, the process of making diagnoses, and the names of medications and psychotherapies used for treatment). It was also hypothesised that the models would be able to differentiate between the support groups, even

after excluding noise topics.

Finally, as the primary aim of this study was to understand which subjective experiences differentiated BPD from the other six disorders with overlapping symptomatology, the topics that best distinguished among the seven corresponding support groups were investigated in detail.

Method

We describe in turn the four main steps of the study method: 1) retrieving posts from the support groups; 2) preprocessing the text of the posts; 3) identifying relevant topics in the posts; and 4) building logistic regression models to identify which topics differentiated between the disorders. Figure 1 depicts an overview of these steps with approximate times for computation for reference. This study was approved by the Macquarie University Human Research Ethics Committee (reference: 52019604512390).

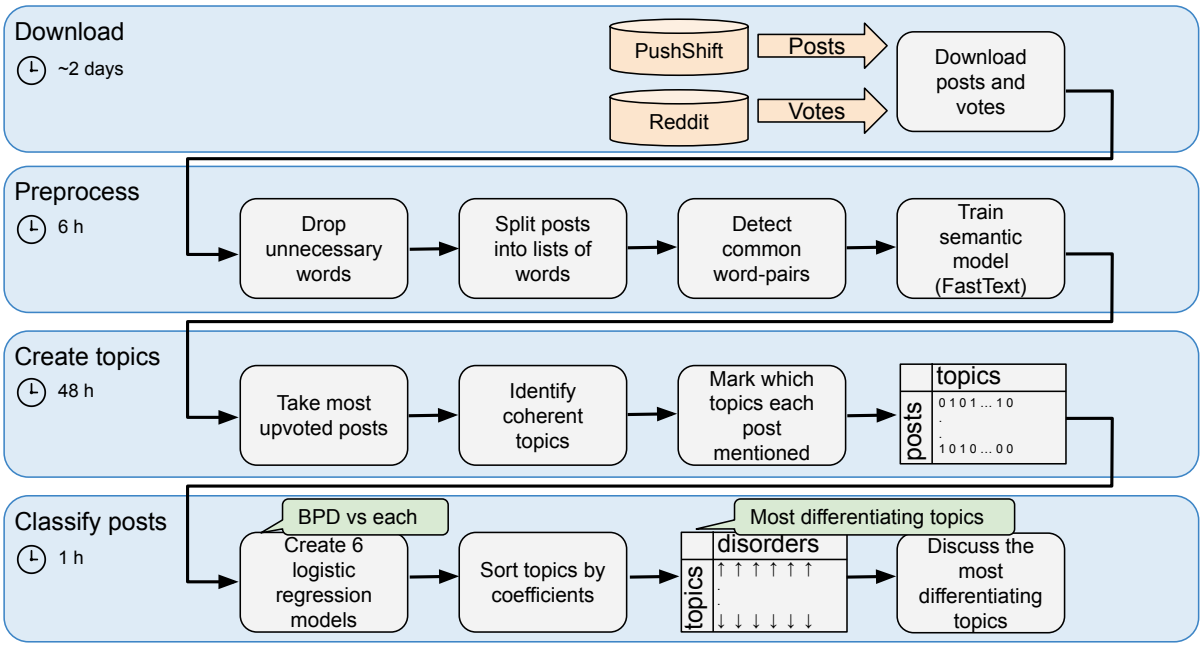

Figure 1. The four major steps of data processing. The net runtimes are measured on a computer with six cores (6*4.8GHz), and 16GB of memory.

Retrieving Posts from Reddit

Content from Reddit was downloaded from PushShift.io, an archive of all Reddit.com posts. The latest count of upvotes for each post was retrieved from Reddit.com using the post's numerical identifier. Upvotes were used in the later stages of processing to identify the most relevant posts in each support group. In line with the ethics approval, only the content of the posts was processed, which contained no personally identifiable information. The download yielded varying numbers of posts from each support group, but the length of the posts tended to be similar across the groups (i.e., all Cohen's $d < 0.2$ compared to BPD). The whole dataset was used for the preprocessing steps (i.e., Preparing the Posts for Modelling and Building the Language Model), and a balanced subset of the most upvoted posts from each support group was used for further analysis, as described below (i.e., Identifying the Topics in the Posts and Data Analysis).

Preparing the Posts for Modelling

Preparation. The titles and bodies of posts were merged into a single block of text. Punctuation and words that contained symbols other than letters, numbers, or underscore characters were removed. These blocks of texts were converted into one-dimensional arrays (i.e., lists of words) to index which words were mentioned in the posts. This is a common technique to process text in machine learning (Hastie, Tibshirani, & Friedman, 2009). All further processing was applied to these one-dimensional arrays.

The number of unique words in the dataset was large (n = 239,945), which made modelling prohibitively complex (referred to as the *curse of dimensionality*; Hastie et al., 2009). Somewhat counterintuitively, classification performance of text-processing *increases* when the number of different words in the dataset (vocabulary size) *decreases*, making

dimensionality reduction both a desirable and routine step in machine learning (Hastie et al., 2009). There are many ways to reduce the number of dimensions, such as excluding items with little meaning or grouping items together. Therefore, following the example of Gkotsis et al. (2017) and Sia, Dalmia, and Mielke (2020), the vocabulary was reduced by eliminating common words (words that carried little meaning, such as "the", "is", "at") and very infrequent words that mostly reflected unusual misspellings (i.e., words that appeared in fewer than 10 posts [0.0008% of posts]). The retained words were then turned into their base forms (*lemmatised*) using the spaCy library for Python (Honnibal & Montani, 2017). This lemmatisation reduces multiple forms of the same word into a single word (e.g., both "runs" and "ran" become "run") to retain as much meaning as possible and further reduce the size of the vocabulary. Using these steps, the total vocabulary size was reduced from 239,945 words to 57,549 words.

Common word-pairs. To improve interpretability and the quality of the topic-models, frequently repeated word-pairs (often called *bigrams*) that behaved like a single word were combined into a single term across the whole dataset using the Normalised Pointwise Mutual Information (NPMI) algorithm from the Gensim software library, with a cutoff score of 0.35 (Lau, Baldwin, & Newman, 2013; Řehůřek & Sojka, 2010). For instance, "downward" and "spiral" were combined into "downward_spiral".

Building the Language Model

The words and word-pairs (henceforth *phrases*) were grouped by their semantic meaning to build cohesive topics. The semantics of the phrases were learned by a word-embedding modelling technique. Word embedding is a data-driven approach that infers the semantics of the phrases in an iterative process. The idea behind this process is that words which repeatedly appear in the same context (i.e., in the same sentence structure with the same

words) tend to have similar meanings — a longstanding and well-validated approach in computational linguistics (Clark, 2015; Firth, 1957; Harris, 1954).

The FastText library was chosen as the embedding framework, as it is resilient to misspellings (Bojanowski et al., 2017). This flexibility was well suited to the common misspellings found in online discussions. Given that the dataset included millions of sentences, instead of using a pre-trained model, a new FastText model was trained on all of the posts from all of the support groups (cf. Grant et al., 2018). When the model finished running, each phrase was mapped into a 100-dimensional space by its semantics. The closer together the two words were, the more semantic meaning they shared (e.g., “happy” and “cheerful” would be close to each other, but far away from “banana” or “density”).

Sampling an Equal Number of Posts From Each Support Group

The number of posts created in the support groups varied substantially over time; therefore, a subsample of posts was drawn from each subreddit to obtain a balanced sample. For each support group, 1,000 equal buckets were created by dividing the duration between the time of the first available post and the last into 1,000 parts (i.e., width = $[date_{last} - date_{first}] / 1000$). All posts within each support group were then allocated to these buckets based on their creation time (i.e., posts in the first time-slice went into the first bucket, and so on). Following that, the two most upvoted posts from each bucket were selected as a representative sample of upvoted posts over time. This process served four purposes. First, the same number of posts were taken from each support group. Second, the sample was large enough to build logistic regression models on; a total of 2,000 posts was shown to be sufficient for this purpose (e.g., Davcheva, 2019). Third, it controlled for the natural growth in posting and voting activity over time (e.g., older posts had a lower vote count but were still selected, along with new posts with high vote counts). Fourth, this

sampling strategy focused on the most meaningful posts in the training set and excluded less relevant posts, as indicated by the upvotes (Kassaeyan, 2016). Overall, the subsampling yielded 2,000 posts from each support group — apart from the antisocial personality group, where all posts were included (n = 861).

Identifying the Topics in the Posts

Based on methodological limitations identified in previous research, the present study set two requirements for identifying semantically coherent clusters (henceforth *topics*) of phrases in the 100-dimensional space. First, the clustering method had to be able to identify similar items but leave ambiguous phrases out. An example would be to create a topic from the phrases “obsess over”, “obsessed”, and “obsess”, and another from “impulsive”, “impulsivity” and “impulsiveness”, while excluding “train” from both groups. Second, the number of topics had to be empirically based, rather than entered manually. Previous research often arbitrarily defined the number of topics (e.g., Thorstad & Wolff, 2019), which can result in unrelated phrases being included in a topic (i.e., the number of topics imposed is not a good fit to the data).

Based on these two requirements for creating semantically coherent topics, the Ordering Points To Identify the Clustering Structure (OPTICS; Ankerst, Breunig, Kriegel, & Sander, 1999) clustering algorithm was chosen. This choice differed from past research, which has often used either k-means clustering — forcefully assigning ambiguous phrases into topics — or Latent Dirichlet Allocation topic modelling — requiring a prespecified number of topics (e.g., Gaur et al., 2018; Sia et al., 2020). OPTICS, on the other hand, met both requirements of this study: 1) it detected varying-density groups, and excluded items from clustering that were too far apart from any clusters (i.e., had ambiguous meaning; Ankerst et al., 1999); 2) it did not require a prespecified number of topics upfront, allowing topics to naturally emerge.

Based on preliminary analyses, a seed-phrase method of creating topics was used that started with a small set of manually selected, representative phrases of underlying topics (Jagarlamudi, Daumé, & Udupa, 2012), yielding 1,744 topics.

An objective naming convention was applied for labelling each group of phrases to limit the bias introduced by subjectively naming topics (Brookes & McEnery, 2019). Specifically, each topic was named with the phrase that was the closest to the centre of the cluster. For instance, if the topic included the phrases "obsess over", "obsessed", and "obsess", and "obsess" was the closest to the mean of these three FastText vectors, then the topic was named "obsess".

These semantically coherent topics, which characterised individuals' subjective experiences of each mental disorder, represented the units of analysis for subsequent steps. The dataset consisted of the subsamples of the most upvoted posts from the seven support groups, indicating whether each post (rows) mentioned each of the 1,744 topics (columns; coded "1" for mentioned and "0" for not mentioned) in seven [posts] * [topics] matrices (one matrix for each of the seven disorder-level support groups).

Data Analysis

The steps so far allowed the last step of the analyses to answer the research question of this study; namely, to identify which subjective experiences differentiated BPD from other disorders. This was achieved in three steps: 1) creating six logistic regression models using topics to predict whether posts belonged to the BPD group or other support groups; 2) using the coefficients from these models to identify which topics were the key differentiators; 3) interpreting the most important topics in detail.

In total, six datasets were created for the logistic regression analyses to compare the BPD [posts] * [topics] matrix with each of the other six support group matrices:

BPD-generalised anxiety disorder, BPD-antisocial personality, BPD-bipolar disorder, BPD-major depressive disorder, BPD-narcissistic personality, and BPD-schizophrenia. Each dataset contained the 2,000 most upvoted posts from the BPD support group, and the 2,000 most upvoted posts from the other support group, as described earlier in the Sampling an Equal Number of Posts From Each Support Group section ($n_{combined}$ = 4,000; the BPD-antisocial pair yielded 2,861 rows due to the lower number of posts in the antisocial personality support group). The outcome variable, an extra column in these combined datasets, indicated whether each post belonged to the BPD (1) or the other disorder support group (0).

Feature selection analyses showed that the top 200 topics — as measured by their chi-square scores on the outcome variable — were sufficient to reach the maximum prediction accuracy across all models. Therefore, for further analysis, only these top 200 topics were used instead of the total 1,744.

Preliminary analyses showed that the data had an adequate predictor-to-outcome ratio, all observations (i.e., posts) were independent of each other, and the correlations between independent variables (i.e., topics) did not suggest high levels of multicollinearity (i.e., all assumptions were met).

Two steps were taken to maximise the generalisability of the results: 1) preventing overfitting of the logistic regression models (i.e., to prevent modelling idiosyncrasies of the training dataset that did not generalise beyond the samples), and 2) measuring how well the models generalised to new, unseen posts.

To prevent overfitting of the models, the logistic regression models used L1 regularisation. L1 regularisation penalises complex modes by forcing small, potentially ambiguous coefficients in the model to be zero, yielding more parsimonious models.

To measure how well the topics generalised beyond present research, the average

classification error on new datasets was estimated, using K-fold cross-validation. The K-fold cross-validation repeated and averaged the cross-validation results 10 times. In each step, a new model was trained on a randomly selected 90% of rows, and the model's AUC score was measured on the excluded 10% (i.e., at each step a new model was trained on n_{train} = 3,600, and measured on n_{test} = 400 rows). AUC evaluates the overall performance of a model; a random classifier would get an AUC score of .5, and a perfect classifier would get a score of 1.0 (Hanley & McNeil, 1982; Hossin & Sulaiman, 2015). In the present study, the K-fold process yielded 10 AUC scores for each support-group pair, and these scores were averaged to get a final, conservative estimate of how well each model generalised to new datasets (henceforth *AUC10FOLD*). If the models received acceptable AUC10FOLD scores (well above .5), then the topics generalised well to new datasets; or, in other words, the subjective experiences differentiated BPD well from the other six disorders.

Identifying key differentiators. The coefficients for each topic from the six logistic regression models were used to identify which subjective experiences were the best differentiators between BPD and the six other disorders. As a first step, the coefficients were placed into a new [topics] * [disorders] matrix: rows represented the topics, columns represented the disorders, and the values within the cells were the logistic regression coefficients (see the first step in Figure 2).

A new sorting method was devised to find clinically relevant experiences. The chosen method was inspired by the Wilcoxon signed-rank test (Wilcoxon, 1945), relying on rank-sums. The main idea was that, while the coefficient values across the logistic regression models were not directly comparable (Mood, 2010), their ranks were. First, the logistic regression coefficients for each topic were ranked within the models (i.e., within each column in the [topics] * [disorders] matrix) between 1 (most important; the highest coefficient) and

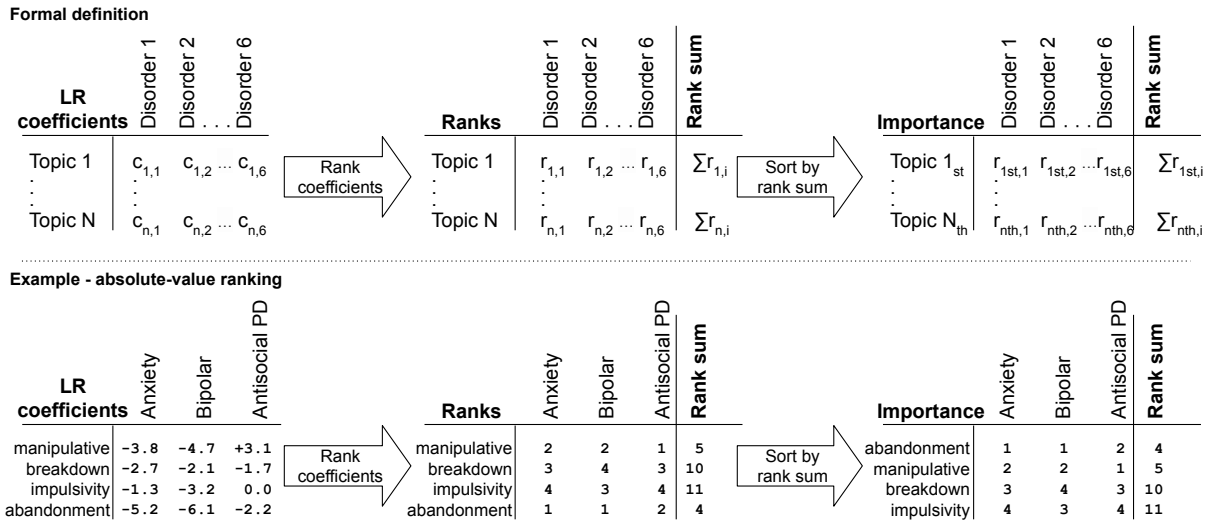

Figure 2. The devised rank-sorting algorithm. Lower ranks imply more discriminative topics. Formal definition on the top, artificial example using absolute-value ranking on the bottom. $c_{x,\,y}$ represents the logistic regression coefficient of $Topic_x$ when predicting $disorder_y$. $r_{x,\,y}$ represents the rank of the $c_{x,\,y}$ coefficient. $r_{ith,\,k}$ represents the rank of the ith most important topic on $disorder_k$. As $\sum r_{1st,\,i} < \sum r_{nth,\,i}$, therefore $Topic_{1st}$ is a better differentiator than $Topic_{nth}$.

200 (least important; the coefficient closest to 0), depicted as the middle step in Figure 2. Each topic (row) therefore had six ranks, each showing how well the topic differentiated between BPD and the other paired disorder. The topic ranks in each row were then summed, creating a new rank-sum column; the lower the rank-sum was, the better the topic was at differentiating between BPD and all other support groups. Therefore, the topic with the smallest rank-sum was the most important differentiator, and so on. For an overview of this sorting method, see Figure 2.

The rank-sums' absolute value was sorted in descending order (bottom example in Figure 2), yielding the most important differential topics, even if the values were conflicting: while all of these subjective experiences were important differentiators, some predicted BPD, and others better predicted another disorder. For instance, while "manipulation" was an important predictor, it described antisocial personality disorder better than it did BPD (i.e., a positive logistic regression coefficient), but it was less typical to anxiety compared to BPD (i.e., a negative coefficient).

Selecting topics for detailed analysis. It was infeasible to discuss all of the 1,744 subjective experiences in the models; therefore, a cutoff for the number of topics was selected. Limiting the number of topics to discuss also increased objectivity and reduced confirmation bias by not allowing cherry-picking favourable items from a large number of data points (Diaconis, 2011). First, to choose which list to discuss the experiences from (raw ranked or absolute-value-ranked list), new logistic regression models were trained by adding topics one by one and plotting the models' AUC10FOLD scores. These preliminary tests showed that topics from the absolute-value-ranked list had consistently higher AUC10FOLD scores than topics from the raw-value rank-sum list. Therefore, the absolute-value-ranked topics were used for further analysis.

The largest changes in AUC10FOLD scores were observed (cf. an elbow plot). The largest AUC10FOLD increase was evident after adding the first few topics to the models, as expected, when the model just started to improve above the .5 baseline, and after adding the 25th topic. Therefore, the first 25 topics were selected for detailed analysis and discussion.

Topic disambiguation. Disambiguation of each of the 25 topics and their constituent phrases was conducted by randomly selecting and reading 30 posts for each topic from the BPD support group (a longstanding method for disambiguation; Mei, Liu, Su, & Zhai, 2006). This step reduced the chance of over-interpreting or misinterpreting the topics by viewing the topic names or phrases out of context — a problem that was evident in previous research. This disambiguation clarified, for instance, whether the topic "manipulative" referred to being manipulative or feeling manipulated. These topics were compared to the DSM-5 criteria for BPD, and where a criterion was not represented in the top 25 topics in the current models, the top 200 topics were examined to identify a match.

Results

Six logistic regression models with the most important 200 differential-topic candidates as predictors were fitted to determine classification performance. Analyses were subsequently repeated after removing noise topics and using the first 25 subjective experiences as predictors (see Table 1 for the AUC10FOLD scores). AUC10FOLD scores reflect the percentage of the BPD posts that received higher values than non-BPD posts. For instance, relying on 25 topics, the model assigned higher values to 64.7% of the BPD posts compared to the bipolar posts (a random classifier would assign 50% to each group).

Table 1
Maximum AUC10FOLD scores across the models

	BD	GAD	MDD	ASPD	NPD	SCZ	Average
Model with 200 topics	.857	.813	.817	.756	.802	.859	.817
noise topics removed	.756	.772	.736	.680	.704	.785	.739
top 25 topics	.647	.653	.657	.529	.629	.665	.630

Note. The top 25 topics were selected from the absolute-value-ranked list, comparing BPD to generalised anxiety disorder (GAD), antisocial personality disorder (ASPD), bipolar disorder (BD), major depressive disorder (MDD), narcissistic personality disorder (NPD), and schizophrenia (SCZ).

Using the subjective experiences as predictors, the models were able to differentiate between the BPD support group and those of the other six disorders better than chance ($AUC10FOLD_{average}$ = .630), relying on only 25 predictors — even after removing noise topics — in line with the hypotheses (i.e., $AUC10FOLD > .5$ in all models). Given these predictors comprise only 1.43% of the total number of topics (1,744), this suggests that these 25 subjective experiences were highly relevant to the groups.

The top 25 topics, their disambiguated meanings, and their correspondence to the DSM-5 criteria of BPD are summarised in Table 2. The top 25 topics were typically more representative of the BPD posts (i.e., negative logistic regression coefficients) but, in some cases, they described other disorders equally well or better than they did BPD (i.e.,

Table 2
The most differentiating subjective experiences between BPD and other disorders

Rank	DSM-5	Topic name (automatically assigned), Description (based on the analysis of 30 random posts)
1	#1	**abandonment rejection**, *Fearing rejection and abandonment*
2	*desc.*	**manipulation**, (<NPD <ASPD) *Perceived as being manipulative*; seeing others as manipulative
3		**breakdown**, Experiencing a crisis (emotional breakdown)
4		**invalidate**, Need for validation; feeling invalidated
5		**ex**, (<NPD) Experiences with past relationships
6	#2	**committed relationship**, (<NPD) *Relationship difficulties (both romantic and otherwise)*
7		**favourite**, Often "favourite person"; referring to interests
8		**yesterday**, (<GAD <BD <MDD <SCZ) Talking about recent, upsetting events
9	#6	**mood fluctuate**, (≈BD) *Rapid mood changes*; weight fluctuation
10	#4	**impulsivity**, (≈ASPD) *Impulsivity, unplanned major decisions*
11	#9	**dissociation derealization**, (≈SCZ) *Feeling detached, not real*
12		**downwards spiral**, (≈MDD) Drifting into a depressive state
13	#6	**oscillate between**, (≈NPD) *Emotional volatility*; instability of self-esteem
14		**behaviour**, (≈ASPD <NPD) Describing generic behaviours
15		**skill**, (≈NPD) Often dialectical behaviour therapy skills, coping skills; general skills
16	*desc.*	**needy clingy**, (≈NPD) *Being needy, overly attached*
17		**jealous**, (≈ASPD <NPD) Being jealous in relationships (both romantic and otherwise)
18		**obsess**, (≈NPD) Excitement about someone or something new; rumination and intrusive thoughts
19		**nuts**, (<BD ≈SCZ) Feeling overwhelmed with intolerable emotions
20	*desc.*	**emotionally abusive**, (≈ASPD) Being emotionally abusive; *feeling abused*
21		**mirror**, (≈NPD) Mirroring others' behaviours; not recognising themselves in the mirror
22	#4	**impulsive reckless**, (≈ASPD) *Similar to impulsivity topic, with "living in the moment" emphasis*
23	*AMPD*	**criticise**, (≈ASPD <NPD) Sensitivity to criticism; *self-criticism*
24	*desc.*	**abuse**, (≈ASPD <NPD) *Childhood abuse or abusive family*; abusing others; substance abuse
25		**anybodys**, (≈GAD) Seeking validation from others ("Does anybody feel like [...]")

Note. Descriptions were based on all the phrases in a topic. Distinct interpretations are separated by ";". The italicised topic descriptions correspond to specific DSM-5 diagnostic criteria, marked with #. *desc.* = in the text-description of BPD in DSM-5, *AMPD* = in Alternative Model for Personality Disorders of DSM-5. GAD = generalised anxiety disorder, ASPD = antisocial personality disorder, BD = bipolar disorder, MDD = major depressive disorder, NPD = narcissistic personality disorder, SCZ = schizophrenia. Based on the logistic regression coefficients, topics were more descriptive of BPD than of the other disorders unless marked with ≈ or <, which denotes equally or more descriptive of another disorder than of BPD, respectively (e.g. ≈NPD means the topic was equally descriptive of NPD, <BD means the topic was more descriptive of BD than of BPD).

coefficients ≥ 0) — a result of picking topics from the absolute-value-ranked list. For instance, the topic "manipulation" described both narcissistic and antisocial personality disorder better than it did BPD, even though it was more typical to the BPD support group than to the other four support groups (major depressive disorder, generalised anxiety disorder, bipolar, and schizophrenia).

Regarding similarities to the DSM-5, seven of the 25 topics directly related to diagnostic criteria of BPD. A further four matched behaviours described in text in the Borderline

Personality Disorder section of the DSM-5, and another related to a dimension in the Alternative Model for Personality Disorders section. Meanwhile, only four topics uniquely identified BPD (abandonment rejection; breakdown; invalidate; favourite), and there were notable similarities between BPD and the other two personality disorders, particularly with narcissistic personality disorder.

Discussion

Given the uncertainties around diagnosing BPD, the present study set out to investigate what subjective experiences — as described in the language of people posting in a BPD support group — could help to diagnose BPD and differentiate it from other disorders. A novel combination of machine learning methods was devised to address previous methodological limitations and identity topics that users discussed in online mental health support groups on Reddit.

Both hypotheses were supported: 1) the models were able to differentiate between the seven different disorder-level support groups based on the subjective experiences described in each post; and 2) this performance was maintained even after removing topics contingent on diagnosis (i.e., noise topics).

Analysis of the Subjective Experiences

Convergent evidence with the DSM-5 and the literature. As mentioned above, the top 25 subjective experiences revealed a strong convergence between the models generated by this study and the DSM-5. The five diagnostic criteria that were reflected in the top 25 subjective experiences were: fear of abandonment (ranked #1), unstable relationships (ranked #6), mood instability (ranked #9), dissociative symptoms (ranked #11), and impulsivity (ranked #22). Notably, fear of abandonment and relationship difficulties were also found to be

key differentiators between BPD and bipolar disorders in previous research, whereas mood instability — a topic shared by BPD and bipolar disorders in the present study — reflected the shared cyclothymic nature of the two disorders (Bayes et al., 2016; Bayes et al., 2019). Furthermore, dissociative symptoms in the present study did not differentiate between schizophrenia and BPD, reinforcing the importance of recognising the differences between the two disorders (e.g., Beatson et al., 2019). Finally, another four topics aligned well with the description of BPD in the main text of the DSM-5 (being manipulative, needy/clingy, emotionally abusive, and abusive childhood).

Apart from the alignment with the DSM-5, several other subjective experiences converged with previous research on the differential diagnosis of BPD. For example, Bayes et al. (2016) also found *sensitivity to criticism* (ranked #23) to be an important differentiator between BPD and other disorders, potentially due to its similarity to rejection sensitivity, which is a core feature of BPD. Regarding *seeking validation* (ranked #4), Linehan (1993) noted it as an important subjective experience of the disorder that needs special attention during therapy. *Downward spiral* (ranked #12) mirrored the definition of affect dysregulation — a criterion of BPD in the DSM-5 — given by Renaud et al. (2012). Importantly, this emotional dysregulation is at the core of BPD, as postulated by the biosocial theory of Linehan (1993), and the semantic differences here are likely due to the present models' capacity to express the experiences in the language of the sufferers. In line with this finding, *obsessive thoughts and behaviours* (ranked #18) is associated with the high arousal often experienced by people with BPD (Linehan, 1993). Finally, *mirror* (ranked #21; interpreted in Table 2 to mean "mirroring others' behaviours") is often mentioned in non-academic sources (e.g., Out of the FOG, n.d.), but it is not typically discussed in the scientific literature and represents a potential addition to the symptoms considered in diagnosing BPD. Overall, these

patterns of convergence in the results suggest that the present study was able to identify valid and clinically relevant distinguishing subjective experiences of BPD in the language of its sufferers, relying on unguided online conversations. These findings could potentially help not only refine how BPD is diagnosed, but also reduce the uncertainty of the diagnosis (frequently reported by practitioners; Sisti et al., 2016), allowing patients to better interpret diagnostic criteria phrased in their own language.

Divergence from the DSM-5. The remaining four criteria from the DSM-5 were not identified in the top 25 subjective experiences, although three of them were reflected in the first 200 topics: suicidal and self-harming behaviour (ranked #46 and #62), anger (ranked #47), and feelings of emptiness (ranked #107). Identity disturbance was not represented in the top 200 topics.). The following sections discuss these four criteria, which these results suggest should perhaps have less weight in the differential diagnosis of BPD.

Self-harming and suicidal behaviour is a well-documented BPD behaviour, and it is estimated that 48-80% of people with BPD engage in deliberate self-harm (Chapman, Specht, & Cellucci, 2005; Gunderson & Hoffman, 2005). However, this statistic can be misleading for diagnostic purposes. First, self-harm and suicidal thoughts are not uncommon in the general population. For example, 15% of surveyed adolescents (age 15-16 years) reported frequent suicidal thoughts (Hawton, 2002), and the lifetime prevalence of self-harm in another study (age 11-18 years) was estimated to be 16-18% (Muehlenkamp, Claes, Havertape, & Plener, 2012). Furthermore, even in a non-clinical adult population (Air Force recruits, M_{age} = 20), the rate of recent self-harm was around 4% (Klonsky, Oltmanns, & Turkheimer, 2003). Taking these base rates into account and using a Bayesian estimate, the predictive power of self-harm on BPD can be as low as 3.2%. Second, self-harm and suicidal thoughts are common in other disorders, such as major depressive disorder, generalised anxiety disorder and bipolar

disorders, especially in the presence of stressors (Clements et al., 2015; Haw, Hawton, Houston, & Townsend, 2001; Hawton, 2002; Klonsky et al., 2003; Skegg, 2005). Accordingly, in our models, self-harm did not differentiate between BPD and bipolar disorder or antisocial personality disorder at all, and was only somewhat more typical to BPD posts than to depression. Overall, in line with previous research, self-harm or suicidal thoughts do not appear to be unique predictors of the experience of BPD.

Anger is another topic that did not differentiate well between BPD and other disorders in the current models: it was equally typical to the other two personality disorders, and only marginally more descriptive of BPD than of schizophrenia. Similarly to self-harm, this could be explained by the prevalence of anger in schizophrenia (Pinkham, Brensinger, Kohler, Gur, & Gur, 2011), and its alignment with *antagonism* in the HiTOP model that is shared across BPD and narcissistic and antisocial personality disorders. This overlap with other personality disorders suggests that anger may not be useful as a specific diagnostic criterion for BPD.

Feelings of emptiness has also been associated with unipolar depression and, to a degree, with anxiety disorders: Klonsky (2008) found robust correlations between feelings of chronic emptiness and unipolar depression, and medium-sized correlation with anxiety disorders. In line with this, feelings of emptiness in the present models did not differentiate between the depression and BPD support groups. Furthermore — similarly to anger — emptiness described both narcissistic and antisocial personality disorder as well as it did BPD, so may be reflecting the antagonism spectrum described in HiTOP. Overall these findings raise questions about the specificity of the feelings of emptiness diagnostic criterion for BPD in DSM-5.

Finally, *identity disturbance* was the only DSM-5 criterion for BPD that was not identified verbatim in the top 200 subjective experiences. However, there were related subjective experiences that referred to similar psychopathology as the DSM-5 criterion, such

as *mirroring* (ranked #21), discussed above, and *oscillate between* (ranked #13, interpreted to mean "instability of self-esteem").

In sum, this study adds to the extant research discussed above that suggests (1) suicidal and self-harming behaviour, (2) anger, (3) feelings of emptiness, and (4) identity disturbance may not differentiate well between BPD and other disorders. It also highlights that mental disorders are typically a combination of many often similar and overlapping behaviours. This symptom-level overlap among disorders has been described in detail elsewhere (e.g., Kotov et al., 2017).

Considerations for new criteria. The present study also identified several subjective experiences that differentiated BPD from the other disorders but are not currently captured by the diagnostic criteria of BPD in the DSM-5. Two in particular might be considered as additions to the diagnostic criteria, due to their salience in everyday events and easy verbal or written administration: 1) *breakdown* (ranked #3) — or, in clinical terms, experiencing a crisis (Campbell et al., 2020) — and 2) *need for validation* (ranked #4), which had been described by Linehan (1993) as a core component of BPD. However, even subtle changes in the diagnostic criteria can result in substantial changes in the prevalence of the disorder (Samuel et al., 2012) and the utility of these characteristics for differential diagnosis would need to be explicitly tested.

Personality disorders in a dimensional framework. As noted above, BPD, narcissistic personality disorder, and antisocial personality disorder appeared to have highly overlapping subjective experiences. Notably, descriptions of manipulation, jealousy, sensitivity to criticism, and childhood abuse were shared across these three disorders. As mentioned above, the similarity between these disorders is well documented in dimensional models of maladaptive personality, in line with the patterns of association found here.

The present study also highlighted additional similarities between subjective descriptions of BPD and narcissistic personality disorder specifically, including relationship difficulties, emotional volatility, neediness, obsession, and mirroring. This, too, is in line with the HiTOP model where antisocial personality disorder is an indicator of both antagonistic and disinhibited externalizing dimensions, whereas narcissistic PD and BPD only indicate antagonism. Finally, three topics uniquely overlapped between BPD and antisocial personality disorder (impulsivity, emotional abuse, and impulse/reckless), potentially capturing the *disinhibition* component in the HiTOP model.

Despite this dimensional overlap, and the similarities between the three personality disorders, in particular, the present results supported the position that BPD has unique subjective experiences that meaningfully distinguish it from other disorders, in line with Campbell et al. (2020). Accordingly, BPD is the only personality disorder with published treatment guidelines from the American Psychiatric Association (2001) or the National Health and Medical Research Council Australia (2013). Furthermore, in moving to a dimensional approach of personality disorders, the International Classification of Diseases 11th Revision kept solely *Borderline Pattern* as a special descriptor, while other personality disorders are described as a combination of different psychopathologies (Bach & First, 2018). In sum, the present results showed robust convergent evidence with both the DSM-5 and the HiTOP model, especially around the structure of personality disorders, reinforcing our confidence in the robustness of the modelling approach.

Practical and Theoretical Implications

Lived experiences of mental disorders. The most important finding of the present study was that the lived experiences of people with BPD could help identify the subjective experiences that differentiate it from other disorders. More broadly, the present results

suggested that diagnostic criteria, such as those in the DSM-5, could be enhanced by taking into account these self-reported life experiences, even if based on online discussion and self-disclosure rather than clinically collected. Well-chosen organically created data — such as that on Reddit — allowed the disorders to be modelled using the language of people with lived experience, strengthened ecological validity, and likely reduced the social desirability bias via anonymity. Accordingly, self-selected support groups appeared to be a strong signal of people with similar experiences and challenges. By using this source of information in a combination of new modelling approaches, the models used here were able to create a ranked list of subjective experiences that distinguished BPD from other related mental disorders. Notably, fear of abandonment emerged as a unique predictor of BPD, in line with the DSM-5 and previous research. The remaining DSM-5 diagnostic criteria, on the other hand, overlapped at least somewhat with other disorders, especially with narcissistic and antisocial personality disorders, giving ways for potential misdiagnosis.

Overall, the similarities between this study's findings and both the DSM-5 and HiTOP suggest these present findings are relevant to both the categorical and the dimensional approaches of mental disorders. This is likely because the identified subjective experiences are meaningful and interpretable in either context. However, as the models aimed to collect the differentiating subjective experiences of BPD, the topics did not constitute a complete description of the experience of the disorder. Further research could investigate modelling a holistic definition of BPD.

Strengths, Limitations, and Future Directions

Signal-to-noise ratio. A particular strength of this study was that it was the first to model subjective experiences of BPD from user-generated content, with the potential to clarify the distinguishing features of this disorder. Importantly, the clarity of the identified

subjective experiences was high, despite the high propensity for noise from using unstructured online posts. Specifically, 84% of the top 25 experiences mapped to symptoms of psychopathology/personality pathology, highlighting a healthy signal-to-noise ratio. Furthermore, as nothing in the technique was specific to BPD or the English language, this novel combination of machine learning methods can also be applied to investigate other disorders and languages as well, potentially identifying cultural nuances to the experience of BPD. Finally, these results did not require expert analysis or any kind of solicited responses from the participants, relying only on the implicit signal of upvotes on the posts to capture what felt important for people discussing BPD. Topics that were difficult to interpret objectively in the context of psychopathology (e.g., "yesterday") do, however, need further consideration if future studies translate these findings into practice.

Self-selection into support groups. An important limitation of the present study was that the people who posted in these support groups did not have to have formal diagnoses of the corresponding disorder. However, as described earlier, diagnosis of BPD is often unreliable (Ng et al., 2019), often given without strictly relying on DSM-5 criteria (e.g., Bayes et al., 2016), and the accuracy of self-diagnosis of mental disorders is in some cases on par with a professional's diagnosis (Lewis, 2017; Stern & Yeomans, 2018). We, therefore, treated self-selection into support groups as an important signal to identify people with similar behaviours and symptomatology. This self-selection could have been problematic if users posted across many support groups (i.e., akin to having multiple diagnoses), but cross-posting across groups in these data was not common (an average of 5.7% users post to multiple groups; Thorstad & Wolff, 2019). Therefore, this overlap was unlikely to have influenced the models substantially. Regardless, the present preliminary findings need replication and validation in clinical settings. Finally, as no demographic or historical data were captured from the

participants, specific demographics might be underrepresented: the support groups on Reddit are used by people who have internet access and are comfortable with using online forums.

Analytical decisions. Finally, while care was taken to justify and measure the impact of the methodological decisions of the present study, ultimately the results are bound by those decisions. For example, no topic-combinations (i.e., interactions) were taken into account in the current models and it should be investigated whether some combinations might serve as better predictors of the disorders(e.g., combining semantically similar topics like ”sadness emptiness” + “empty hollow” could further reduce the dimensionality and increase the interpretability of the models).

Conclusions

The present study used a novel modelling approach to identify the subjective experiences that characterise BPD. The results showed convergent evidence with the DSM-5, the HiTOP model, and previous research on differential diagnosis. Results also indicated that some DSM-5 criteria may not be effective for diagnosing BPD: self-harm, anger, and feelings of emptiness were common across multiple disorders; relationship difficulties, impulsivity, and emotional volatility were shared with antisocial and narcissistic personality disorders as well. However, fear of abandonment emerged as a key differentiator of BPD; experiencing a breakdown and the need for validation were also good candidates for experiences that uniquely identify BPD. There were also many subjective experiences of BPD that were discussed in the support group that are not covered here, and may represent important features for a comprehensive and holistic description of BPD. Overall, this study indicates that applying machine learning methods to user-generated content could enhance our understanding of BPD by incorporating subjective experiences of the disorder — expressed in the language of individuals with lived experience — contributing to a more accurate and

timely diagnosis, and reducing the suffering of people with BPD.

Authorship

A.H. developed the study concept, performed the data collection, analysis and interpretation, and wrote the original draft of the paper. M.F. and M.D. supervised the project, contributed to the study design and interpretation of the data, and provided critical revisions of the manuscript. All authors approved the final version of the paper for submission.

References

American Psychiatric Association. (1980). *Diagnostic and Statistical Manual of Mental Disorders: DSM-III* (3rd ed). Washington, D.C.: American Psychiatric Association.

American Psychiatric Association. (2001). *Practice Guideline for the Treatment of Patients with Borderline Personality Disorder*. Washington, D.C: American Psychiatric Association.

American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders: DSM-5* (5th ed). Arlington, VA: American Psychiatric Association.

Ankam, V. (2016). *Big Data Analytics*. Packt Publishing. Retrieved July 1, 2020, from https://public.ebookcentral.proquest.com/choice/publicfullrecord.aspx?p=4699930

Ankerst, M., Breunig, M. M., Kriegel, H.-P., & Sander, J. (1999). Optics: Ordering Points to Identify the Clustering Structure. *ACM SIGMOD Record*, *28*(2), 49–60. doi:10.1145/304181.304187

Australian Bureau of Statistics. (2019, December). Australian Demographic Statistics. Retrieved from https://www.abs.gov.au/AUSSTATS/abs@.nsf/mf/3101.0

Bach, B., & First, M. B. (2018). Application of the ICD-11 Classification of Personality Disorders. *BMC Psychiatry*, *18*(1), 351. doi:10.1186/s12888-018-1908-3

Bayes, A., McClure, G., Fletcher, K., Román Ruiz del Moral, Y. E., Hadzi-Pavlovic, D., Stevenson, J. L., … Parker, G. (2016). Differentiating the Bipolar Disorders from Borderline Personality Disorder. *Acta Psychiatrica Scandinavica*, *133*(3), 187–195. doi:10.1111/acps.12509

Bayes, A., Parker, G., & Paris, J. (2019). Differential Diagnosis of Bipolar II Disorder and Borderline Personality Disorder. *Current Psychiatry Reports*, *21*(12), 125. doi:10.1007/s11920-019-1120-2

Beatson, J. A., Broadbear, J. H., Duncan, C., Bourton, D., & Rao, S. (2019). Avoiding Misdiagnosis When Auditory Verbal Hallucinations Are Present in Borderline Personality Disorder: *The Journal of Nervous and Mental Disease*, *207*(12), 1048–1055. doi:10.1097/NMD.0000000000001073

Beattie, D., Murphy, S., Burke, J., O'Connor, H., & Jamieson, S. (2019). Service User Experiences of Clinical Psychology Within an Adult Mental Health Service: An Ipa Study. *Mental Health Review Journal*, *24*(3), 171–182. doi:10.1108/MHRJ-02-2018-0005

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017, June 19). Enriching Word Vectors with Subword Information. arXiv: 1607.04606 [cs]. Retrieved February 17, 2020, from http://arxiv.org/abs/1607.04606

Brookes, G., & McEnery, T. (2019). The Utility of Topic Modelling for Discourse Studies: A Critical Evaluation. *Discourse Studies*, *21*(1), 3–21. doi:10.1177/1461445618814032

Campbell, K., Clarke, K.-A., Massey, D., & Lakeman, R. (2020). Borderline Personality Disorder: To Diagnose or Not to Diagnose? That Is the Question. *International Journal of Mental Health Nursing*, inm.12737. doi:10.1111/inm.12737

Chakravorti, D., Law, K., Gemmell, J., & Raicu, D. (2018). Detecting and Characterizing Trends in Online Mental Health Discussions. *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*, 697–706. doi:10.1109/ICDMW.2018.00107

Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., & Blei, D. (2009, January). Reading Tea Leaves: How Humans Interpret Topic Models. (Vol. 32, pp. 288–296).

Chapman, A. L., Specht, M. W., & Cellucci, T. (2005). Borderline Personality Disorder and Deliberate Self-Harm: Does Experiential Avoidance Play a Role? *Suicide and Life-Threatening Behavior*, *35*(4), 388–399. doi:10.1521/suli.2005.35.4.388

Choudhury, M. D., & De, S. (2014). Mental Health Discourse on Reddit: Self-Disclosure, Social Support, and Anonymity. *Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014*, 10.

Clark, S. (2015). Vector Space Models of Lexical Meaning. *Handbook of Contemporary Semantics*. doi:10.1002/9781118882139.ch16

Clements, C., Jones, S., Morriss, R., Peters, S., Cooper, J., While, D., & Kapur, N. (2015). Self-Harm in Bipolar Disorder: Findings from a Prospective Clinical Database. *Journal of Affective Disorders*, *173*, 113–119. doi:10.1016/j.jad.2014.10.012

Cristea, I., Gentili, C., Cotet, C., Palomba, D., Barbui, C., & Cuijpers, P. (2017). Efficacy of Psychotherapies for Borderline Personality Disorder: A Systematic Review and Meta-analysis. *JAMA Psychiatry*, *74*(4), 319. doi:10.1001/jamapsychiatry.2016.4287

Davcheva, E. (2019, December). Classifying Mental Health Conditions Via Symptom Identification: A Novel Deep Learning Approach.

DeShong, H. L., Mullins-Sweatt, S. N., Miller, J. D., Widiger, T. A., & Lynam, D. R. (2016). Development of a Short Form of the Five-Factor Borderline Inventory. *Assessment*, *23*(3), 342–352. doi:10.1177/1073191115581475

Diaconis, P. (2011, October 17). Theories of Data Analysis: From Magical Thinking Through Classical Statistics. In D. C. Hoaglin, F. Mosteller, & J. W. Tukey (Eds.), *Wiley Series in Probability and Statistics* (pp. 1–36). doi:10.1002/9781118150702.ch1

Ernala, S. K., Birnbaum, M. L., Candan, K. A., Rizvi, A. F., Sterling, W. A., Kane, J. M., & De Choudhury, M. (2019). Methodological Gaps in Predicting Mental Health States from Social Media: Triangulating Diagnostic Signals. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19* (pp. 1–16). The 2019 CHI Conference. doi:10.1145/3290605.3300364

Firth, J. R. (1957). A Synopsis of Linguistic Theory. *1952-59*, 1–32.

Gaur, M., Kursuncu, U., Alambo, A., Sheth, A., Daniulaityte, R., Thirunarayan, K., & Pathak, J. (2018). "Let Me Tell You About Your Mental Health!": Contextualized Classification of Reddit Posts to DSM-5 for Web-based Intervention. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management - CIKM '18* (pp. 753–762). The 27th ACM International Conference. doi:10.1145/3269206.3271732

Gkotsis, G., Oellrich, A., Velupillai, S., Liakata, M., Hubbard, T. J. P., Dobson, R. J. B., & Dutta, R. (2017). Characterisation of Mental Health Conditions in Social Media Using Informed Deep Learning. *Scientific Reports*, *7*(1), 45141. doi:10.1038/srep45141

Grant, R., Kucher, D., Leon, A., Gemmell, J., & Raicu, D. (2017, November). Discovery of Informal Topics from Post Traumatic Stress Disorder Forums. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)* (pp. 452–461). 2017 IEEE International Conference on Data Mining Workshops (ICDMW). doi:10.1109/ICDMW.2017.65

Grant, R., Kucher, D., Leon, A., Gemmell, J., Raicu, D., & Fodeh, S. (2018). Automatic Extraction of Informal Topics from Online Suicidal Ideation. *BMC Bioinformatics*, *19*(S8), 211. doi:10.1186/s12859-018-2197-z

Guimaraes, A., Balalau, O., Terolli, E., & Weikum, G. (2019). Analyzing the Traits and Anomalies of Political Discussions on Reddit. *Proceedings of the International AAAI Conference on Web and Social Media*, *13*(01), 205–213. Retrieved from https://www.aaai.org/ojs/index.php/ICWSM/article/view/3222

Gunderson, J. G., & Hoffman, P. D. (Eds.). (2005). *Understanding and Treating Borderline Personality Disorder: A Guide for Professionals and Families* (1st ed). Washington, DC: American Psychiatric Pub.

Hanley, J., & McNeil, B. (1982). The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve. *Radiology*, *143*(1), 29–36. doi:10.1148/radiology.143.1.7063747

Harris, Z. (1954). Distributional Structure. WORD, *10*(2-3), 146–162. doi:10.1080/00437956.1954.11659520

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed). New York, NY: Springer.

Haw, C., Hawton, K., Houston, K., & Townsend, E. (2001). Psychiatric and Personality Disorders in Deliberate Self-Harm Patients. *British Journal of Psychiatry*, *178*(1), 48–54. doi:10.1192/bjp.178.1.48

Hawton, K. (2002). Deliberate Self Harm in Adolescents: Self Report Survey in Schools in England. *BMJ*, *325*(7374), 1207–1211. doi:10.1136/bmj.325.7374.1207

Honnibal, M., & Montani, I. (2017). *Spacy 2: Natural Language Understanding with Bloom Embeddings, Convolutional Neural Networks and Incremental Parsing*. Retrieved from https://spacy.io/

Hossin, M., & Sulaiman, M. (2015). A Review on Evaluation Metrics for Data Classification Evaluations. *International Journal of Data Mining & Knowledge Management Process*, *5*(2), 01–11. doi:10.5121/ijdkp.2015.5201

Hyler, S. E., Rieder, R. O., Williams, J. B. W., Spitzer, R. L., Hendler, J., & Lyons, M. (1988). The Personality Diagnostic Questionnaire: Development and Preliminary Results. *Journal of Personality Disorders*, *2*(3), 229–237. doi:10.1521/pedi.1988.2.3.229

Jagarlamudi, J., Daumé, H., & Udupa, R. (2012). Incorporating Lexical Priors into Topic Models. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 204–213). USA: Association for Computational Linguistics.

Kassaeyan, K. (2016). *Factors Affecting Upvoting Intention on Social Bookmarking Sites* (Luleå University of Technology, Business Administration and Industrial Engineering).

Klonsky, E. D. (2008). What is Emptiness? Clarifying the 7th Criterion for Borderline Personality Disorder. *Journal of Personality Disorders*, *22*(4), 418–426. doi:10.1521/pedi.2008.22.4.418

Klonsky, E. D., Oltmanns, T. F., & Turkheimer, E. (2003). Deliberate Self-Harm in a Nonclinical Population: Prevalence and Psychological Correlates. *American Journal of Psychiatry*, *160*(8), 1501–1508. doi:10.1176/appi.ajp.160.8.1501

Kotov, R., Krueger, R. F., Watson, D., Achenbach, T. M., Althoff, R. R., Bagby, R. M., … Zimmerman, M. (2017). The Hierarchical Taxonomy of Psychopathology (HiTOP): A dimensional alternative to traditional nosologies. *Journal of Abnormal Psychology*, *126*(4), 454–477. doi:10.1037/abn0000258

Lantz, B. (2013). The Large Sample Size Fallacy. *Scandinavian Journal of Caring Sciences*, *27*(2), 487–492. doi:10.1111/j.1471-6712.2012.01052.x

Lau, J. H., Baldwin, T., & Newman, D. (2013). On collocations and topic models. *ACM Transactions on Speech and Language Processing*, *10*(3), 1–14. doi:10.1145/2483969.2483972

Leichsenring, F. (1999). Development and First Results of the Borderline Personality Inventory: A Self-Report Instrument for Assessing Borderline Personality Organization. *Journal of Personality Assessment*, *73*(1), 45–63. doi:10.1207/S15327752JPA730104

Leichsenring, F., Leibing, E., Kruse, J., New, A., & Leweke, F. (2011). Borderline Personality Disorder. *The Lancet*, *377*(9759), 74–84. doi:10.1016/S0140-6736(10)61422-5

Leontieva, L., & Gregory, R. (2013). Characteristics of Patients with Borderline Personality Disorder in a State Psychiatric Hospital. *Journal of Personality Disorders*, *27*(2), 222–232. doi:10.1521/pedi_2013_27_078

Lewis, L. F. (2017). A Mixed Methods Study of Barriers to Formal Diagnosis of Autism Spectrum Disorder in Adults. *Journal of Autism and Developmental Disorders*, *47*(8), 2410–2424. doi:10.1007/s10803-017-3168-3

Lilienfeld, S., & Lynn, S. J. (2014, May 22). Errors/Biases in Clinical Decision Making. In R. Cautin & S. Lilienfeld (Eds.), *The Encyclopedia of Clinical Psychology* (pp. 1–9). doi:10.1002/9781118625392.wbecp567

Linehan, M. (1993). *Cognitive-Behavioral Treatment of Borderline Personality Disorder*. New York: Guilford Press.

Livesley, W. J. (2020). Why Is an Evidence-Based Classification of Personality Disorder so Elusive? *Personality and Mental Health*, pmh.1471. doi:10.1002/pmh.1471

McCrae, R. R., & Costa, P. T. (2003). *Personality in Adulthood: A Five-Factor Theory Perspective* (2nd ed). New York: Guilford Press.

Mei, Q., Liu, C., Su, H., & Zhai, C. (2006). A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In *Proceedings of the 15th international conference on World Wide Web - WWW '06* (p. 533). The 15th international conference. doi:10.1145/1135777.1135857

Molnar, C. (2020). *Interpretable machine learning ; a guide for making black box models explainable*.

Mood, C. (2010). Logistic Regression: Why We Cannot Do What We Think We Can Do, and What We Can Do About It. *European Sociological Review*, *26*(1), 67–82. doi:10.1093/esr/jcp006

Muehlenkamp, J. J., Claes, L., Havertape, L., & Plener, P. L. (2012). International Prevalence of Adolescent Non-Suicidal Self-Injury and Deliberate Self-Harm. *Child and Adolescent Psychiatry and Mental Health*, *6*(1), 10. doi:10.1186/1753-2000-6-10

National Health and Medical Research Council Australia. (2013). *Clinical Practice Guideline for the Management of Borderline Personality Disorder*. Canberra: National Health and Medical Research Council.

Ng, F., Townsend, M., Miller, C., Jewell, M., & Grenyer, B. (2019). The Lived Experience of Recovery in Borderline Personality Disorder: A Qualitative Study. *Borderline Personality Disorder and Emotion Dysregulation*, *6*(1), 10. doi:10.1186/s40479-019-0107-2

Out of the FOG. (n.d.). Top 100 Traits of People Who Suffer from Personality Disorders. Retrieved July 9, 2020, from https://outofthefog.website/top-100-trait-blog/2015/11/4/mirroring

Paris, J. (2010). Estimating the Prevalence of Personality Disorders in the Community. *Journal of Personality Disorders*, *24*(4), 405–411. doi:10.1521/pedi.2010.24.4.405

Pinkham, A. E., Brensinger, C., Kohler, C., Gur, R. E., & Gur, R. C. (2011). Actively paranoid patients with schizophrenia over attribute anger to neutral faces. *Schizophrenia Research*, *125*(2-3), 174–178. doi:10.1016/j.schres.2010.11.006

Řehůřek, R., & Sojka, P. (2010, May 22). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 workshop on new challenges for NLP frameworks* (pp. 45–50). Valletta, Malta: ELRA.

Renaud, S., Corbalan, F., & Beaulieu, S. (2012). Differential Diagnosis of Bipolar Affective Disorder Type Ii and Borderline Personality Disorder: Analysis of the Affective Dimension. *Comprehensive Psychiatry*, *53*(7), 952–961. doi:10.1016/j.comppsych.2012.03.004

Samuel, D. B., Miller, J. D., Widiger, T. A., Lynam, D. R., Pilkonis, P. A., & Ball, S. A. (2012). Conceptual Changes to the Definition of Borderline Personality Disorder Proposed for DSM-5. *Journal of Abnormal Psychology*, *121*(2), 467–476. doi:10.1037/a0025285

Samuel, D. B., Sanislow, C. A., Hopwood, C. J., Shea, M. T., Skodol, A. E., Morey, L. C., … Grilo, C. M. (2013). Convergent and Incremental Predictive Validity of Clinician, Self-Report, and Structured Interview Diagnoses for Personality Disorders Over 5 Years. *Journal of Consulting and Clinical Psychology*, *81*(4), 650–659. doi:10.1037/a0032813

Shatte, A., Hutchinson, D. M., & Teague, S. J. (2019). Machine Learning in Mental Health: A Scoping Review of Methods and Applications. *Psychological Medicine*, *49*(09), 1426–1448. doi:10.1017/S0033291719000151

Sia, S., Dalmia, A., & Mielke, S. J. (2020). Tired of Topic Models? Clusters of Pretrained Word Embeddings Make for Fast and Good Topics Too!

Siddaway, A., Quinlivan, L., Kapur, N., O'Connor, R. C., & de Beurs, D. (2020). Cautions, Concerns, and Future Directions for Using Machine Learning in Relation to Mental Health Problems and Clinical and Forensic Risks: A Brief Comment on "Model Complexity Improves the Prediction of Nonsuicidal Self-Injury" (fox Et Al., 2019). *Journal of Consulting and Clinical Psychology*, *88*(4), 384–387. doi:10.1037/ccp0000485

Sisti, D., Segal, A., Siegel, A., Johnson, R., & Gunderson, J. (2016). Diagnosing, Disclosing, and Documenting Borderline Personality Disorder: A Survey of Psychiatrists' Practices. *Journal of Personality Disorders*, *30*(6), 848–856. doi:10.1521/pedi_2015_29_228

Skegg, K. (2005). Self-harm. *The Lancet*, *366*(9495), 1471–1483. doi:10.1016/S0140-6736(05)67600-3

Stern, B. L., & Yeomans, F. (2018). The Psychodynamic Treatment of Borderline Personality Disorder. *Psychiatric Clinics of North America*, *41*(2), 207–223. doi:10.1016/j.psc.2018.01.012

Thorstad, R., & Wolff, P. (2019). Predicting Future Mental Illness from Social Media: A Big-Data Approach. *Behavior Research Methods*, *51*(4), 1586–1600. doi:10.3758/s13428-019-01235-z

Warrender, D. (2015). Staff Nurse Perceptions of the Impact of Mentalization-Based Therapy Skills Training When Working with Borderline Personality Disorder in Acute Mental Health: A Qualitative Study: Staff Perceptions of Mbt-S for Bpd in Acute Mental Health. *Journal of Psychiatric and Mental Health Nursing*, *22*(8), 623–633. doi:10.1111/jpm.12248

Wilcoxon, F. (1945). Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, *1*(6), 80. doi:10.2307/3001968. JSTOR: 10.2307/3001968

Yin, Z., Sulieman, L. M., & Malin, B. A. (2019). A Systematic Literature Review of Machine Learning in Online Personal Health Data. *Journal of the American Medical Informatics Association*, *26*(6), 561–576. doi:10.1093/jamia/ocz009