How to identify anxiety on Twitter: a natural language processing approach.

Daniel Zarate[1], Vasileios Stavropoulos[1], Maria Prokofieva[1], and Bruno Schivinski[2]

[1]Institute for Health and Sport, Victoria University, Melbourne, Australia

[2]School of Media and Communication, RMIT, Melbourne, Australia

Corresponding author: Vasileios Stavropoulos; vasileios.stavropoulos@vu.edu.au

Abstract

Background: Social media posts have been suggested as a promising/novel source of mental health information, when examined via natural language processing means. To realize this opportunity, distinct Twitter post/language patterns related to reported anxiety diagnosis were explored. **Methods:** A number of 233.000 tweets made by 605 users (300 reporting anxiety diagnosis and 305 not) over six months were comparatively analysed, considering user behaviour, linguistic enquiry word count (LIWC), and sentiment analysis. **Results:** Supervised machine learning models (Random Forests, Naïve Bayes, and LASSO Regression) showed a high prediction power, with the Random Forests model accurately distinguishing 84.5% of anxiety reporting and/or not users (0.93 AUC). Additionally, a Latent Profile Analysis (LPA) identified four posting profiles characterized by high sentiment, low sentiment, self-distancing, and normative behaviour. **Conclusion:** The digital footprint of anxiety on one's Twitter posts presented with high frequency of words conveying negative sentiment, low frequency of positive sentiment, reduced frequency of posting, and lengthier texts. These distinct patterns enabled highly accurate prediction of reported anxiety diagnosis. On this basis, appropriately resourced, awareness raising, online mental health campaigns are advocated.

Keywords: Anxiety; Digital footprint; Cyber-Phenotype; Machine learning; Natural language processing; Sentiment Analysis; Text analysis; Twitter

Social media engagement as a source of mental health information

The highly popular usage of social media allows a person to openly express their opinions, ideas, feelings and thoughts to others, via their instantly accessible personal devices (i. e. smartphones; computers etc; Zarate et al., 2022). Interestingly, the language chosen in one's social media posts has been featured as a valuable source of mental health information (Insel, 2018). Notably, the term '*cyberphenotype'* has been introduced to describe how online behaviour, including social media posts, may indirectly/unintentionally operate as a health/mental-health footprint (Zarate et al., 2022). Such diagnostic potential, especially considering a user's mental health, is maximized by the constant (*moment-to-moment)* and naturalistic/ non-research induced (*in-situ*) flow of information/data, passively (i.e. without the user's conscious involvement) captured (Zarate et al., 2022). These qualities are viewed to outweigh the validity and reliability offered by traditional clinical practice methods, projecting the decoding of mental health information entailed in social-media participation as a an imperative priority (Cuthbert & Insel, 2013; Insel, 2018).

Indeed, social networking sites (SNS) provide researchers with an ideal source of information, procuring access to an individuals' uncensored voices/thoughts and narratives (Ngai et al., 2015). This could be particularly important for assessing one's mental health, as the language adopted by a user to express themselves online may be less contained/pre-planned (Insel, 2018). Such qualities are reinforced by a sense of anonymity and distance (pseudo-anonymity) that encourages the release of rich and meaningful information that may be otherwise remain elusive/restricted (Insel, 2018). Furthermore, over time observations of how a user conducts themselves on SNS (i. e. time, frequency and content of posts/ interactions with others over lengthier periods of time; > 6 months), often portray/ project their relationships-narratives as well as their identity-narratives (i.e. How they view/experience their engagement with others; How they view/engage with their own self;

Denzin, 2004; McAdams, 2010; Stavropoulos et al., 2022). In other words, SNS posts' self-expressions could either consciously (purposefully) and/or unconsciously (latently) encapsulate content/statements that confirm/align with deeper/core self-appraisals/conceptions/beliefs, likely associated to the user's mental health (Denzin, 2004; Graci et al., 2018).

In that context, many studies to date have employed multiple social networking sites (e.g., Twitter, Reddit, Weibo, Facebook) to assess linguistic expressions and accurately detect a range of psychological disorders, including depression, suicide, schizophrenia, eating disorders, anxiety, etc. (Chancellor & De Choudhury, 2020; Coppersmith et al., 2014; Coppersmith et al., 2015). For example, researchers have identified significant associations between discernible language patterns (including intonation, word rate, fluency, grammatical form, and lexical selection) and mood (Larsen et al., 2020), symptoms of depression (Gkotsis et al., 2017) and psychosocial stressors (Mowery et al., 2017). However, while a vast proportion of such studies focused on assessing depression or suicide, only a minority of those emphasized on anxiety, inviting for further related research (Dutta et al., 2018; Ireland & Iserman, 2018; Saifullah et al., 2021; Shen & Rudzicz, 2017). To address these recommendations and to expand the available knowledge, the current project will focus on identifying anxiety on twitter posts, while adopting a novel and advanced analyses' methodology.

Identifying anxiety online

Anxiety presentations involve worry and apprehension about one or more different conditions or stimuli, often marked by bodily symptoms of physical tension (e. g. accelerated breath and heartbeat; Taschereau-Dumouchel et al., 2022). Literature suggests that anxiety symptoms/ behaviours are normally distributed in the community from minimum to

maximum (i. e. all individuals are anxious at varying rates, with a minority experiencing extremely high anxiety; ElHafeez et al., 2022). Although, objective/healthy/non-problematic anxiety constitutes a proportionate/realistic reaction to a perceived threat; excessive/disproportionate anxiety, to the extent that one's wellbeing is compromised, is regarded as the common denominator across several debilitating mental-health conditions such as generalized anxiety (i.e. symptoms occur across a variety of life domains), social anxiety (i.e. symptoms relate to how the person could be perceived by others) , specific phobias (i.e. symptoms evolve around a specific object/condition), or panic attacks (i.e. symptoms are accompanied by episodes of elevated/acute fear and physical discomfort entailing palpitations, sweating etc.; American Psychiatric Association, 2013).

Subsequently, higher anxiety has been evidenced to interfere with one's cognitions and behaviours (i.e. risk evaluation thoughts; risk avoidance actions), as well as the language they choose/select to express themselves, particularly when online (Sonnenschein, Hofmann, Ziegelmayer, & Lutz, 2018; Settanni and Marengo, 2015). For instance, more anxious individuals have been shown to verbally/linguistically express with higher negative affect, lower positive affect, increased self-criticism, lower self-efficacy expectations, experiential avoidance and tensed utterance, whilst uncensored emotional social media expressions are more common among younger users (Berman et al., 2010; Joiner & Blalock, 1995; Rook et al., 2022; Settanni and Marengo, 2015; Smith & Jones, 2013; Sonnenschein et al., 2018; Woodgate et al., 2020).

In that line, studies exploring digital traces of mental health disorders have repeatedly emphasized social media posts to identify anxiety (Zarate et al., 2022). To address this aim, a sequence of steps/stages involving (a) accessing (b) analyzing and (c) predicting anxiety-related information has been consistently followed, although with significant variations considering the analysis methods suggested (Chancellor & De Choudhury, 2020). Firstly, the

creation of large user/content databases frequently requires the use of application programming interfaces (APIs) to retrieve and organize information/data in meaningful ways (i. e. relevant concepts/words and users/profiles associated to anxiety are accessed and collated; Zarate et al., 2022). Secondly, the dataset(s) composed are examined via natural language processing (NLP) techniques, which aim to quantify meaningful signals/patterns of anxiety within a given corpus of text (i.e., compilation of texts; anxious patterns of expressions; Chowdhary, 2020). For example, NLP techniques may quantify/assess lexical and semantic forms of anxiety in a text via the inclusion of *n*-gram analysis (i. e. assessing anxiety linked continuous successions of words, symbols or prefixes-tokens), language inquiry word count (LIWC; e. g. proportion of words falling under different linguistic, psychological and topical categories tied to anxiety; comparing and matching anxious and non-anxious language styles; systematically extracting anxiety related meaning), and sentiment analysis (i.e., identification of anxious emotional tone in a specific text), among others (Guntuku et al., 2017). To maximize the prediction power of such text analysis' findings, recent studies have additionally employed machine learning algorithms (i. e. analysis methods that "learn" to progressively improve their accuracy via leveraging on data accumulation and testing; Singh et al., 2016; Zarate et al., 2022). These enrich text-analysis based, anxiety prediction models making them more accurate (Singh et al., 2016).

Not surprisingly, previous studies analysing social media texts/posts to detect anxiety demonstrated promising accuracy and invited further research in the area (Coppersmith et al., 2014; Coppersmith et al., 2015; Ive et al., 2018; Ireland & Iserman, 2018; Shen et al., 2017). For example, Ireland and Iserman (2018) used a decision tree algorithm to detect anxiety-related posts on Reddit with a 92% accuracy. Similarly, Ive et al. (2018) used hierarchical neural models to detect anxiety-related posts with 82% accuracy. Finally, Coppersmith et al. (2015), used Twitter posts and character *n*-gram language models to differentially

diagnose/identify anxiety from 9 other distinct mental health presentations. Despite such important steps, to the best of the authors' knowledge, no studies to date have aimed to more accurately describe the cyber-phenotype of anxiety via concurrently considering/examining sentiment analysis, user-behavior (i.e., frequency, text length, and time of posting), and LIWC evidence. The combination of such methods could inform anxiety classification and *profiling* models (i. e. models where anxiety associated social media content typologies could be simultaneously portrayed on the basis of all these features, such that anxious and non-anxious individuals are correctly classified). Indeed, several studies to date have successfully used latent profile analyses informed by psychometric indicators (i. e. questionnaire scores) to distinguish different types of digital media users, such as gamers (Billieux et al., 2015; Kovacs et al., 2022). The advantage of following such an approach would enable, aside of distinctly connecting different social media text and usage patterns with anxiety (i.e. variable focused research), to more holistically and accurately portray anxious types/profiles of social media users (i.e. person focused research; Stavropoulos, Motti-Stefanidi and Griffiths, 2022).

The present study

Aspired by this opportunity, and to expand the available knowledge, the present study innovatively co-examined posts of twitter users reporting and/or not anxiety diagnosis over a period of six months, aiming to: (i) decode/identify a series of linguistic twitter expressions/patterns, entailing user behaviour, LIWC and sentiment analysis, significantly associated to reporting anxiety diagnosis and; (ii) to collectively consider such patterns, as latent profile analysis indicators, to accurately describe different twitter posting typologies and their links to anxiety. These person/profiled focused and methodologically innovative aims were additionally enhanced by the comparative employment of several machine algorithms to maximize prediction power. Accordingly, the findings are expected to have significant practical contributions. Firstly, from a clinical perspective, the early and cost-

efficient identification of people who suffer from anxiety represents the potential to optimize treatment outcomes through their timely engagement. Secondly, knowledge of one's anxious cyber-phenotype could help tailor minimally invasive, resource-saving, online interventions to address anxiety in the broader community. Thus, the following research questions were proposed:

Research Question 1: Can an individual's anxious/non-anxious status be accurately predicted by their linguistic expressions on Twitter?

Research Question 2: What is the number and nature of profiles that best describe the sample considering their linguistic expression on Twitter, and how do these profiles (if at all) relate to one's reported anxiety diagnosis?

Method

Data collection

Data collection commenced after obtaining approval from the Victoria University Research Ethics Committee (HRE21-114). Following the method employed by Coppersmith et al. (2014, 2015), the rtweet package for RStudio (Kearney, 2019) was used to access the most recent six months of Twitter activity from a self-diagnosed group of 300 Twitter users who publicly posted the phrase "I have been diagnosed with anxiety". Similarly, a control group of Twitter users who did not post such phrase on their timeline was sourced (n=305). All data used in this study has been publicly posted on Twitter between May and November 2021 and made available through Twitter's API. The collections of tweets used here include a total of 233.000 tweets in English and do not include direct messages, retweets, or data marked as private by the author.

Data processing

Data processing involved a series of steps to quantify user behaviour and linguistic expression on Twitter, including user activity, language inquiry word count (LIWC), and sentiment analysis. Following the method employed by Silge and Robinson (2022), Tidytext (Silge & Robinson, 2016), tidyverse (Wickham et al., 2019), and lubridate RStudio packages (Grolemund & Wickham, 2011) were used to compare characteristics of linguistic expression on Twitter between the 'anxious' and the 'non-anxious' groups. Average Tweet length and time of posting (including time of day and day of the week) across anxious/non-anxious groups were compared. Specifically, tweets were (a) compiled into a 'corpus', (b) *tokenized* (separate tweets into single words), and (c) stop words were removed (stop words are frequent words such as "the", "is", "of", and may not add value to lexical analyses). A bag-of-words approach was then used to quantify frequency of terms using the *term frequency*inverse document frequency* function (*tf idf*; Silge & Robinson, 2022). The *tf-idf* approach is commonly employed in NLP to normalize term frequency across documents (in this case, tweets) and thus obtain a score of 'term-salience', with higher scores representing higher importance. Finally, *sentiment analysis* was conducted to classify and quantify emotional intent in tweets. Bing (Liu, 2015), nrc (Mohammad, 2021), and afinn (Nielsen, 2011) lexicons/dictionaries were used to classify words into sentiment categories and ascribe emotional valence to tweets. The nrc lexicon classifies 13875 words into ten different sentiments including anger, anticipation, disgust, fear, joy, negative, positive, sadness, surprise and trust. The bing lexicon classifies 6786 words into positive/negative sentiment, and the afinn lexicon ascribes emotional value to 2477 words ranging from -3 to +3.

Prediction and classification models

Prediction and classification models were created to identify the presence/absence of anxiety and identify unique cyberphenotypical characteristics in Twitter users. Specifically, to answer *RQ1*, the tidymodels RStudio package (Kuhn & Wickham, 2020) was used to build predictive models (null, Naïve Bayes, LASSO regression, and Random Forests) using eight model indicators (text length, tf-idf, hour of the day, day of the week, weekday/weekend, sentiment bing, sentiment, nrc, sentiment afinn) and thus identify whether users belonged to the 'anxious' or 'non-anxious' group. A ten-fold stratified cross-validation resampling method was used to train predictive models, and assessment of model accuracy involved examination confusion matrix and area under the curve (AUC) with higher AUC indices representing higher accuracy of class prediction.

Finally, to answer *RQ2*, the tidyLPA RStudio package (Rosenberg et al., 2019) was used to identify latent profiles (Latent Profile Analysis, LPA). A maximum likelihood estimator (MLE) was used to estimate parameterization combinations, whilst information/classification criteria (Akaike information criterion [AIC], Bayesian information criterion [BIC] and standardized entropy [*h*]) were used to determine the optimum number of latent profiles (see Hvitfeldt & Silge [2022] for an explanation of machine learning models, and Kovacs et al. [2022] and Masyn [2013] for an explanation of LPA). *Figure 1* presents the data flow process from data acquisition to prediction and classification models.

- *Figure 1* -

Results

Twitter user behaviour and sentiment analysis

A series of statistical tests were conducted in RStudio (2020) to determine potential differences in user behavior across anxious/non-anxious groups. 'Anxious' users tweeted less

frequently (χ^2[1]=6465, p<.001 add effect size if easy) and posted longer tweets than the non-anxious (*Kruskal-Wallis* $\chi^2_{[1]}$=47050, p<.001; $M_{anxious}$=92.9 / $M_{non\text{-}anxious}$=51 add effect size if easy). Additionally, 'anxious' users tweeted evenly throughout the week and more frequently during 4pm to 11pm ($F_{[1,44]}$=18.67, p<.001; effect size if easy; 47% of tweets between these times), whereas the 'non-anxious' users tweeted more frequently on Sundays ($\chi^2_{[6]}$=693, p<.001; effect size if easy) and evenly spread throughout the day. See Supplementary Table 1 for a description of tweets grouped by 'anxious' and 'non-anxious' users, including tweet frequency, length, time of the day, day of the week, and text valence.

Sentiment analysis revealed significant differences across groups. Specifically, the bing lexicon determined that 'anxious' users posted negative words more frequently (56% negative, 44% positive) than the 'non-anxious' users (33% negative, 67% positive; $\chi^2_{[1]}$=12107, p<.001; effect size). Similarly, the nrc lexicon determined that 'anxious' users Tweeted terms expressing negative emotions more frequently, whereas the 'non-anxious' users used positive emotions more frequently. For example, the 'anxious' group used anger-related words (e.g., abandon, abhorrent) 9.5% of the time compared to 5.8% in the 'non-anxious' users, and fear-related words (e.g., absence, badness) 10.4% of the time compared to 5.3% in the 'non-anxious' users (see Supplementary Table 2 for a comprehensive list of the most frequently used sentiment-related words by group). Finally, the sentiment distribution (i.e., emotional valence ascribed to tweets using the afinn lexicon) was significantly different across groups (*Kolmogorov-Smirnov D*=0.26, p<.001) with 'anxious' users showing lower sentiment value ($M_{sentiment}$=-0.30) compared to the 'non-anxious' users ($M_{sentiment}$=1.00). Supplementary *Figure 1* presents an illustration of Twitter user behavior and sentiment analysis discriminated by 'anxious' and 'non-anxious' users.

Classification and profiling models

To answer *Research Question 1*, supervised machine learning models were fitted to classify Twitter users into the 'anxious' and 'non-anxious' users. Specifically, eight model indicators (day of the week, weekday/weekend, hour of the day, bing-lexicon, nrc-lexicon, afinn-lexicon, text length, and tf-idf) were used to fit four machine learning models (null, LASSO-regression, naïve Bayes, and Random Forests). The null model was used to assess randomness of classification without model indicators, and as expected, it correctly 'guessed' every second observation (51% accuracy, 0.5 AUC) showing appropriate randomness of selection. As seen in Table 1 and Figure 2 (left panel), the Random Forests model produced the highest classification accuracy (84.5%, 0.93 AUC) followed by the naïve Bayes (77.7%, 0.84 AUC) and the LASSO-regression (64.3%, AUC 0.86). Specifically, the Random Forests model correctly identified 9 out of 10 true positive cases ('anxious' Twitter users correctly identified as part of the anxious group) compared to 8 out of 10 true negative cases ('non-anxious' Twitter users correctly identified as part of the 'non-anxious' users; Supplementary *Figure 2*). This is not unexpected considering that the model indicators were devised to assess the cyberphenotype of anxiety, and unobserved variables could affect the effectiveness of correctly classifying Twitter users in the 'non-anxious' group. A permutation-based test of importance identified that text length, followed by positive/negative words (bing lexicon), and text emotional valence (afinn lexicon) were the most important variables to maintain accuracy of classification using the Random Forests approach (*Figure 2*, right panel).

- Table 1-

- Figure 2 -

To answer *Research Question 2*, a Latent Profile Analysis (LPA) was used to identify the optimum number of user profile considering their linguistic expression on Twitter. Of the

possible variance-covariance parameter combination, only the class-invariant diagonal parameterization (CIDP) model with equal variances and covariances fixed to zero converged on a solution. Specifically, this model assumes equal variability in model indicators for all latent profiles (equal variance) and no relationships across different profiles (covariance fixed to zero; for an explanation of possible variance-covariance parameterization models see Kovacs et al., 2022). As seen in *Figure 3* (left panel), increasing the number of latent profiles resulted in decreased model errors (AIC and BIC). However, the CIDP model including four latent profiles showed appropriate AIC, BIC, and the highest profile heterogeneity (*standardized entropy h* = 0.93) and was therefore selected as optimum fit (Supplementary Table 3 presents AIC, BIC, entropy, N-min and bootstrapped likelihood ratio test for model 1-CIDP with two to six latent profiles). Accordingly, the share of Twitter users in each latent profile were 40.8% in Profile 1 (*n*=210), 25.2% in Profile 2 (*n*=130), 32% in Profile 3 (*n*=165), and 2% in Profile 4 (*n*=10). Table 2 displays standardized mean scores discriminated by Twitter user latent profile.

- Table 2/Figure 3 -

Twitter user latent profiles were described considering mean standardized user behavior, sentiment expressed in tweets, and frequency of salient terms. As seen in *Figure 3*-right panel, the four latent profiles showed different characteristics. Specifically, Profile 1 was characterized by low sentiment value (-1SD sentiment bing and afinn, and -0.73SD sentiment nrc). Profile 2 showed high sentiment value (+1.3SD sentiment bing and afinn, and +1SD sentiment nrc). Profile 3 was defined by average model indicator values. Finally, Profile 4 was characterized by +2.7SD tf-idf. In this profile, individuals used the words ‘you’ (*n*=583, 7%) and ‘everyone’ (*n*=0, 0%) less frequently compared to the rest of the sample (*n*=303855, 16%; *n*=28565, 1.5%), thus named as the “self-distancing” tweeter users. Can we discuss this

more? I don't understand it. Supplementary *Figure 3* presents a comparison of the five most frequently used words between profile 4 and the rest of the sample.

You need to add a small chi-square result section here to show how the membership of different twitter posting profiles related to being diagnosed or not with anxiety. It is expected. This is your main argument considering how your study advances the literature.

Discussion

This study sought to decode the cyber-phenotype of anxiety using linguistic expressions in social media and to identify different latent profiles of users suffering from anxiety. To address these aims, a natural language processing approach was used to identify user behavior and patterns of linguistic content using a corpus of available tweets purposely accessed via the Twitter API. Overall, Twitter users with a self-disclosed diagnosis of anxiety Tweeted less frequently, posted longer tweets, and used language conveying negative sentiments more frequently than those without a self-disclosed diagnosis of anxiety. Additionally, the Random Forests model correctly predicted 84.5% of users' group membership ('anxious'/'non-anxious'). Finally, four distinct profiles of Twitter users were identified, describing users who expressed high sentiment (41% of users), low sentiment (25%), normative (32%), and self-distancing (2%) language. How do these profiles compare in relation to a self-disclosed anxiety diagnosis. Taken together, these results represent important implications for cost-efficiently and accurately identifying anxiety symptoms expressed on social media platforms.

Identifying anxiety in Twitter posts

Considering the first study aim/research question, significant differences in linguistic expression and user behavior between Twitter users with a self-disclosed diagnosis of anxiety ('anxious') and those without such self-disclosed diagnosis ('non-anxious') were evident.

Specifically, self-disclosed ‘anxious’ users tweeted less frequently and posted longer tweets than non-self-disclosed “anxious’ users. In line with these results, Dutta et al. (2018) observed a reduction in social interactions through online platforms between ‘anxious’ users and their strong online ties/connections, indicating a fear of negative evaluations. Berman et al. (2010) suggest that exaggerated beliefs about being evaluated negatively represent cognitive distortions (from a cognitive behavioral therapy [CBT] framework perspective) and may lead to experiential avoidance (from an acceptance and commitment therapy [ACT] perspective, ACT. In this context, ‘anxious’ Twitter users may show reduced social interaction due to learned maladaptive internal responses (e.g., inferiority, self-criticism, lack of self-compassion) that maintain and reproduce unhelpful patterns of behavior, giving rise to anxiety and fear of scrutiny, thus restricting their interactions with others via posts on twitter (Wright et al., 2017)

Interestingly, the results observed here indicate that while Twitter users with self-disclosed diagnosis of anxiety tweeted less frequently, they posted lengthier texts than the ‘non-anxious’ group. This highlights the possibility of a dichotomous cognitive process in which anxious individuals either avoid posting on social media for fear of negative evaluations or may reversely post lengthier texts due to being overly concerned with minimizing errors and perfectionism (Gregersen & Horwitz, 2002; Ong & Twohig, 2022). Ong and Twohig (2022) proposed that when worried, some people tend to excessively think about future communication mistakes, aiming to prevent them via overly-elaborated and thus lengthier messages/statements. This is reinforced by literature suggesting that anxiety-induced cognitive biases may generate a “black and white” perspective of the world, eventuating perfectionistic engagement with their surrounding (Wright et al., 2017). However, considering the limited available empirical evidence supporting this interpretation, further studies may aim to expand on this area.

Another important difference between Twitter users with a self-disclosed diagnosis of anxiety and the control group resided in the sentiment valence embedded in their tweets. Interestingly, the words most frequently used were similar for both groups, with *good*, *love*, and *happy* being the most frequent positive words and *bad* and *hate* being the most frequent negative words. However, terms reflecting negative affect were more commonly posted by 'anxious' users across all three lexicons employed here (see Supplementary Tables 1 and 2 and Supplementary *Figure 1*). In line with Woodgate et al. (2020), this observation suggests that anxiety-affected individuals may frequently communicate their worry, lack of confidence, negative self-image, and emotional dysregulation.

Prediction and classification of anxiety in Twitter users

The above discussed differences regarding the cyber-phenotypical characteristics between Twitter users with a self-disclosed diagnosis of anxiety and 'non-anxious' users enabled accurate group membership classifications., when enhanced via machine learning techniques. All tested supervised machine learning models showed good classification accuracy, with Random Forests correctly identifying 9 out of 10 'anxious' users and 8 out of 10 'non-anxious' users (84.5% accuracy, 0.93 AUC; see *Figure 2*). The most reliable variables for correct classification of group membership using the Random Forests model were text length and sentiment analyses, indicating that these variables should be considered for prediction/assessment of anxiety via linguistic expressions on social media platforms. Overall, and irrespective of the classification algorithm employed, the results suggest that the chosen model indicators provide sufficient information to accurately detect and predict symptoms of anxiety.

In that line, a latent profile analysis, on the basis of the indicators assessed (i.e. user behaviour, LIWC and sentiment analysis) suggested that four distinct profiles denoted salient

latent cyberphenotypical characteristics of Twitter users in this sample. Specifically, 40.8% of users in this sample showed low sentiment valence in their linguistic expressions via Twitter (-1SD; *Figure 3* right panel). Most users in the low sentiment profile (66%) disclosed a diagnosis of anxiety on Twitter, confirming that 'anxious' users are more likely to use language that conveys negative affect (including anger, fear, disgust, and sadness; Mohammad, 2021; Woodgate et al., 2020). Additionally, 32% and 25% of users in this sample showed normative (mean values) and high sentiment valence (+1SD) in their linguistic expressions respectively. Interestingly, 31% of users who disclosed a diagnosis of anxiety on Twitter also showed high sentiment valence in their posts. This suggests that the cyberphenotype of anxiety is co-informed by other elements beyond low valence of texts (e.g., text length, frequency and time of posting, etc.), and a combination of these elements should be incorporated in models predicting anxiety based on an individual's social media activity. Finally, 2% of Twitter users posted the word *you* significantly less and were thus categorized as the self-distancing group This needs more explanation; why less "you" could imply self-distancing. Previous research suggested that self-focused attention, often due to experienced anxiety/distress, may result in self-distancing practices (i.e., diminished social interaction) and is associated with negative affect (Mor & Winquist, 2002). However, individuals showing self-distancing language did not exhibit above-average negative emotional valence or increased use of the first-person pronoun suggesting the existence of an interesting combination of elements. Thus, one could assume that the fact that twitter-users identified here as anxious, were those who publicly posted their anxiety diagnosis status may interfere with these findings. Specifically, the rather not elevated negative affect reported by those classified as belonging in the fourth profile, could reflect their level of acceptance and embracing of their condition, to the extent that they felt comfortable enough to announce it

online. Considering the above sample limitations, further research may seek to explore this interpretation.

Conclusions, implications and limitations

These results significantly contribute to understanding anxiety through one's linguistic expression on social media platforms, such as Twitter. Specifically, and in line with past literature, findings highlight that individuals suffering from anxiety may use language showing negative affect and reduced positive affect, often entailing statements related to lack of confidence, fear, and worry (Woodgate et al., 2020). Furthermore, findings suggested a reduced frequency of posts from those disclosing and anxiety diagnosis. Interestingly, individuals enduring anxiety symptoms may likely engage in experiential avoidance and fear of negative evaluations, increasing their self-distancing practices, such as posting on Twitter, while reducing their motivation for social interactions (Berman et al., 2010). This is particularly important considering the positive impact either offline (including social activities, hobbies, outdoor activities, sports, etc.) and online (e. g. social media communities) social support networks have on mental health (Li et al., 2021). Following that line, the significantly lengthier posts of those disclosing an anxiety diagnosis could indicate their excessive concerns regarding minimizing errors and presenting as perfect as possible (Ong & Twohig, 2022). Indeed, results suggested that the machine learning prediction models informed by the above differences have promising results regarding the opportunity of automating reliable anxiety assessment on the basis of an individual's twitter activity. Finally, four distinct posting profiles were revealed, with those belonging in… and in being more likely to have self-reported an anxiety diagnosis. Such information can further enhance anxiety assessment capacities by emphasizing more holistically on one's texting patterns (i.e. user behaviour, LIWC and sentiment analysis).

These findings bear significant epidemiological, assessment, prevention and intervention implications. Firstly, from an epidemiological perspective, using highly naturalistic methods such as cyber-phenotyping would facilitate more accurate estimation of prevalence and incidence rates. Specifically, considering that individuals suffering from anxiety might feel averse to voicing their psychological ailments due to stigma and lack of awareness, current statistics might not accurately represent the prevalence rates of such disorders. Secondly, from a clinical assessment perspective, the efficient and cost-effective identification of people who suffer mental health issues via their social media posting activity represents the potential to optimize treatment outcomes through their timely preventive engagement. For example, deploying mental health campaigns dedicated to predicting and detecting anxiety may facilitate accessing relevant information and resources to develop understanding and awareness, promoting action to address such presentations. Finally, from an intervention perspective, the knowledge of one's digital phenotype based on Twitter use could help tailor personalized applications to their recipients' profiles, maximizing the effectiveness of interventions. For example, one's cyberphenotype profile may contribute to efficiently guiding the required intervention strategy (i. e. what works for whom approach).Taken together, these results are expected to provide the basis for devising pioneering services designed to help individuals at risk of suffering from anxiety and potentially other comorbid psychological issues.

Despite these important contributions, the results reported here need to be interpreted in the context of significant limitations. Firstly, considering this study used self-disclosed diagnoses of anxiety, the allocation of group membership did not follow rigorous clinical assessments, and thus results should be interpreted with caution. Secondly, the 'non-anxious' group comprised Twitter users who did not post the phrase "I have been diagnosed with anxiety". Therefore, this method can only allow speculation of 'healthy' mental health, and it

could be feasible that individuals suffering from anxiety do not disclose such a diagnosis on Twitter. Thirdly, the current sample represents a small proportion of Twitter users, and thus results presented here may need to be validated with larger samples derived from other social media sources. Fourthly, further studies may consider how (if at all) sentiment valence in social media posts changes at different milestones from receiving a psychiatric diagnosis.

References

American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders* (5th Ed.). Author. https://doi.org/10.1176/appi.books.9780890425596

Berman, N. C., Wheaton, M. G., McGrath, P., & Abramowitz, J. S. (2010). Predicting anxiety: the role of experiential avoidance and anxiety sensitivity. *Journal of Anxiety Disorders, 24*, 109-113. https://doi.org/10.1016/j.janxdis.2009.09.005

Billieux, J., Thorens, G., Khazaal, Y., Zullino, D., Achab., S. & Van del Linden, M. (2015). Problematic involvement in online gamers: a cluster analytic approach. *Computers in Human Behavior, 43*, 242-250. https://doi.org/10.1016/j.chb.2014.10.055

Chancellor, S., & De Choudhury, M. (2020). Methods in predictive techniques for mental health status on social media: a critical review. *npj Digital Medicine, 3*(43). https://doi.org/10.1038/s41746-020-0233-7

Chowdhary, K. R. (2020). *Fundamentals of Artificial Intelligence*. Springer. https://www.springer.com/gp/book/9788132239703

Coppersmith, G., Dredze, M., & Harman, C. (2014). Quantifying mental health signals in Twitter. *Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signals to Clinical Reality*, pp. 51-60. https://aclanthology.org/W14-3207.pdf

Coppersmith, G., Dredze, M., Harman, C., & Hollingshead, K. (2015). From ADHD to SAD: analyzing the language of mental health on Twitter through self-reported diagnoses. *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical*

Psychology: From Linguistic Signals to Clinical Reality, pp. 1-10. https://aclanthology.org/W15-1201/

Cuthbert, B. N., & Insel, T. R. (2013). Toward the future of psychiatric diagnosis: the seven pillars of RDoC. *BMC Medicine, 11*(126), 1-8. https://doi.org/10.1186/1741-7015-11-126

Denzin, N. (2004). Symbolic interactionism. In L. Pranee (Ed.), *Handbook of Research Methods in Health Social Sciences* (pp. 151-172). Singapore: Springer.

Dutta, S., Ma, J., & De Choudhury, M. (2018). Measuring the impact of anxiety on online social interactions. *Proceedings of the Twelfth International AAAI Conference on Web and Social Media (ICWSM 2018)*, pp. 584-587. https://ojs.aaai.org/index.php/ICWSM/article/view/15081

Gkotsis, G., Oellrich, A., Velupillai, S., Liakata, M., Hubbard, T. J. P., Dobson, R. J. B., & Dutta, R. (2017). Characterization of mental health conditions in social media using informed deep learning. *Scientific Reports, 7*(45141). https://doi.org/10.1038/srep45141

Gregersen, T., & Horwitz, E. K. (2002). Language learning and perfectionism: anxious and non-anxious language learners' reactions to their own oral performance. *The Modern Language Journal, 86*(4), 562-570. https://doi.org/10.1111/1540-4781.00161

Grolemund, G., & Wickham, H. (2011). Dates and times made easy with lubridate. *Journal of Statistical Software, 40*(3), 1-25. https://www.jstatsoft.org/v40/i03/

Guntuku, S. C., Yaden, D. B., Kern, M. L., Ungar, L. H., Eichstaedt, J. C. (2017). Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences, 18*, 43-49. http://dx.doi.org/10.1016/j.cobeha.2017.07.005

Hvitfeldt, E., & Silge, J. (2022). *Supervised machine learning for text analysis in R*. CRC Press. https://smltar.com/

Insel, T. R. (2018). Digital phenotyping: a global tool for psychiatry. *World Psychiatry, 17*(3), 276. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6127813/

Ireland, M. E., & Iserman, M. (2018). Within and between-person differences in language used across anxiety support and neutral reddit communities. *Proceedings of the Fifth*

Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic, pp. 182-193. https://aclanthology.org/W18-0620.pdf

Ive, J., Gkotsis, G., Dutta, R., Stewart, R., & Velupillai, S. (2018). Hierarchical neural model with attention mechanisms for the classification of social media text related to mental health. *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pp. 69-77. https://aclanthology.org/W18-0607/

Joiner, T. E. Jr., & Blalock, J. A. (1995). Gender differences in depression: the role of anxiety and generalized negative affect. *Sex Roles, 33*(1-2), 91-108. https://doi.org/10.1007/BF01547937

Kearney, M. W. (2019). Rtweet: collecting and analyzing Twitter data. *Journal of Open Source Software, 4*(42), 1829. https://doi.org/10.21105/joss.01829

Kovacs, J., Zarate, D., de Sena Collier, G., Tran, T. T. D., & Stavropoulos, V. (2022). Disordered gaming: the role of a gamer's distress profile. *Canadian Journal of Behavioural Science*. Advance online publication. http://dx.doi.org/10.1037/cbs0000335

Kuhn, M., & Wickham, H. (2020). *Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles* [Computer software]. https://www.tidymodels.org

Larsen, M. E., Boonstra, T. W., Batterham, P. J., O'Dea, B., Paris, C. & Christensen, H. (2020). We feel: mapping emotion on Twitter. *IEEE Journal of Biomedical Health Information, 19*(4), 1246-1252. https://doi.org/10.1109/JBHI.2015.2403839

Li, F., Luo, S., Mu, W., Li, Y., Ye, L., Zheng, X., ... & Chen, X. (2021). Effects of sources of social support and resilience on the mental health of different age groups during the COVID-19 pandemic. *BMC psychiatry, 21*(1), 1-14.

Liu, B. (2015). *Sentiment analysis: mining sentiments, opinions, and emotions* (1st Ed.). Cambridge University Press.

Masyn, K. E. (2013). Latent class analysis and finite mixture modeling: statistical analysis. In T. D. Little (Ed.), *The Oxford handbook of quantitative methods* (pp. 551-611). Oxford University Press.

McAdams, D. P. (2010). Narrative Identity. In S. J. Schwartz, K. Luyckx, & V. L. Vignoles (Eds.), *Handbook of identity theory and research* (pp. 99-116). Springer. https://link.springer.com/content/pdf/10.1007/978-1-4419-7988-9.pdf

Mohammad, S., M. (2021). Sentiment analysis: automatically detecting valence, emotions and other affectual states from text. *Emotion Measurement, 2.* https://doi.org/10.48550/arXiv.2005.11882

Molina, S., Borkovec, T. D., Peasley, C., & Person, D. (1998). Content analysis of worrisome streams of consciousness in anxious and dysphoric participants. *Cognitive Therapy and Research, 22*(2), 109-123. https://doi.org/10.1023/A:1018772104808

Mor, N., & Winquist, J. (2002). Self-focused attention and negative affect: A meta-analysis. *Psychological Bulletin, 128*(4), 638–662. https://doi.org/10.1037/0033-2909.128.4.638

Mowery, D., Smith, H., Cheney, T., Stoddard, H., Coppersmith, G., Bryan, C., & Conway, M. (2017). Understanding depressive symptoms and psychosocial stressors on Twitter: a corpus-based study. *Journal of Medical Internet Research, 19*(2), e48. https://doi.org/10.2196/jmir.6895

Ngai, E. W. T., Tao, S. S. C., & Moon, K. K. L. (2015). Social media research: theories, constructs and conceptual frameworks. *International Journal of Information Management, 35*(1), 33-44. https://doi.org/10.1016/j.ijinfomgt.2014.09.004

Nielsen, F. A. (2011). A new ANEW: evaluation of a word list of sentiment analysis in microblogs. *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages,* pp. 93-98. http://arxiv.org/abs/1103.2903

Ong, C. W., & Twohig, M. P. (2022). *The Anxious Perfectionist: How to Manage Perfectionism-Driven Anxiety Using Acceptance and Commitment Therapy*. New Harbinger Publications.

Rook, L., Mazza, M. C., Lefter, J., & Brazier, F. (2022). Toward linguistic recognition of generalized anxiety disorder. *Frontiers in Digital Health,* 4:779039. https://doi.org/10.3389/fdgth.2022.779039

Rosenberg, J., Beymer, P., Anderson, D. van Lissa, C., & Schmidt, J. (2019). tidyLPA: an R package to easily carry out latent profile analysis (LPA) using opoen-source or commercial software. *Journal of Open Source Software, 3*(30), 978. https://doi.org/10.21105/joss.00978

RStudio Team (2020). *RStudio: Integrated Development for R*. RStudio. http://www.rstudio.com/.

Saifullah, S., Fauziah, Y., & Aribowo, A. S. (2021). Comparison of machine learning for sentiment analysis in detecting anxiety based on social media data. *Journal Informatika, 15*(10, 45-55. https://doi.org/10.26555/jifo.v15i1.a20111

Settanni, M., & Marengo, D. (2015). Sharing feelings online: studying emotional well-being via automated text analysis of Facebook posts. *Frontiers in psychology*, *6*, 1045.

Shen, J. H., & Rudzicz, F. (2017). Detecting anxiety on Reddit. *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology,* pp. 58-65. https://aclanthology.org/W17-3107.pdf

Sonnenschein, A. R., Hofmann, S. G., Ziegelmayer, T., & Lutz, W. (2018). Linguistic analysis of patients with mood and anxiety disorders during cognitive behavioral therapy. *Cognitive behaviour therapy*, *47*(4), 315-327.

Silge, J. & Robinson, D. (2016). Tidytext: text mining and analysis using tidy data principles in R. *Journal of Open Source Software, 1*(3). https://dx.doi.org/10.21105/joss.00037

Silge, J. & Robinson, D. (2022). *Text Mining with R: A Tidy Approach.* O'Reilly. https://www.tidytextmining.com/index.html

Singh, A., Thakur, N., & Sharma, A. (2016). A review of supervised machine learning algorithms. *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, pp. 1310-1315. https://ieeexplore.ieee.org/abstract/document/7724478

Smith, A. R., & Jones, D. M. (2013). *Handbook of Human Performance Volume 3: State and Trait.* Academic Press.

Spielberger, C. D. (1966). *Anxiety and Behavior.* Academic Press. https://doi.org/10.1016/C2013-0-12378-1

Stavropoulos, V., Motti-Stefanidi, F., & Griffiths, M. D. (2021). Risks and opportunities for youth in the digital era: A cyber-developmental approach to mental health. *European Psychologist*.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., Francois, R., Grolemund, G., Hayes, A., Herny, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Muller, K., Ooms, J., Robinson, D., Seidel, D. P., …& Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software, 4*(43), 1686. https://doi.org/10.21105/joss.01686

Woodgate, R. L., Tailor, K., Tennent, P., Wener, P., & Altman, G. (2020). The experience of the self in Canadian youth living with anxiety: a qualitative study. *PLoS ONE, 15*(1), e0228193. https://doi.org/10.1371/journal.pone.0228193

Wright, J. H., Brown, G. K., Thase, M. E., & Ramirez Basco, M. (2017). *Learning cognitive-behavior therapy: An illustrated guide* (2nd Ed.). American Psychiatric Association Publishing.

Zarate, D., Stavropoulos, V., Ball, M., de Sena Collier, G., & Jacobson, N. C. (2022). Exploring the digital footprint of depression: a PRISMA systematic literature review of the empirical evidence. *BMC Psychiatry, 22*(1), 421. https://doi.org/10.1186/s12888-022-04013-y

Table 1. Machine learning models

Model	Tunning	Accuracy	AUC	TP	TN	FP	FN
Null	N.A.	0.510	0.500	33	34	32	33
LASSO-regression	λ 0.014	0.643	0.862	21	62	5	41
Naïve Bayes	N.A.	0.777	0.838	13.8	16.2	3.5	5.1
Random Forests	Mtry = 2, Min_n = 25	0.845	0.930	56	53	6	14

Note: Accuracy = number of correct classification divided by total observations. AUC = The area under the curve ranges from 0 to 1 with higher values indicating higher model performance. TP = True positives. TN = True negatives. FP = False positives. FN = False negatives. λ = Penalty term applied to LASSO regressions. Mtry = Number of random variables for recursive partitioning employed at each split (i.e., decision) to minimize node impurity; Min n = Minimum number of data points in a node required for the node to be split further.

Table 2. Description of latent profiles including standardized scores of model indicators

Profile	Text Length	Day of the week	Weekday / Weekend	Hour	Sent Bing	Sent NRC	Sent AFINN	tf-idf
Low sentiment (40.8%)	0.16	0.07	0.05	0.29	-0.97	-0.73	-0.98	0.07
High sentiment (25.2%)	-0.03	-0.19	-0.18	-0.39	1.36	1.02	1.34	-0.16
Normative (32%)	-0.19	0.06	0.04	-0.05	0.14	0.10	0.16	-0.13
Self-distancing (2%)	0.25	0.04	0.38	-0.38	0.39	0.44	0.43	2.68

Note. This table shows standardized (z scores) values for each model indicator discriminated by latent profiles. The 'Low sentiment' and 'High sentiment' profiles are characterized by ~1SD below/above the mean in word sentiment respectively. The 'Self-distancing' class is characterized by high tf-idf denoting a different word use distribution (i.e., the type of words used).

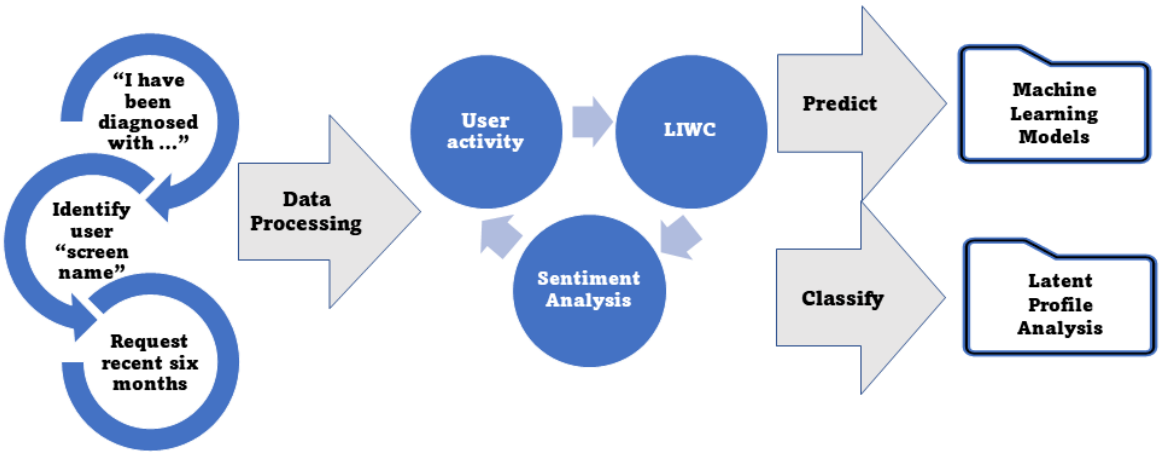


Figure 1. Here we see the data flow process including data retrieval, data processing, and prediction/classification. The first step involved accessing a user database through the Twitter API. The second step involved data processing to obtain model indicators representing user activity (frequency, text length, and time of posts), sentiment analysis, and language enquiry word count (LIWC). Finally, model indicators were used to predict (using Naïve Bayes, Random Forests, and LASSO regression) and classify (using Latent Profile Analysis) 'anxious' users.

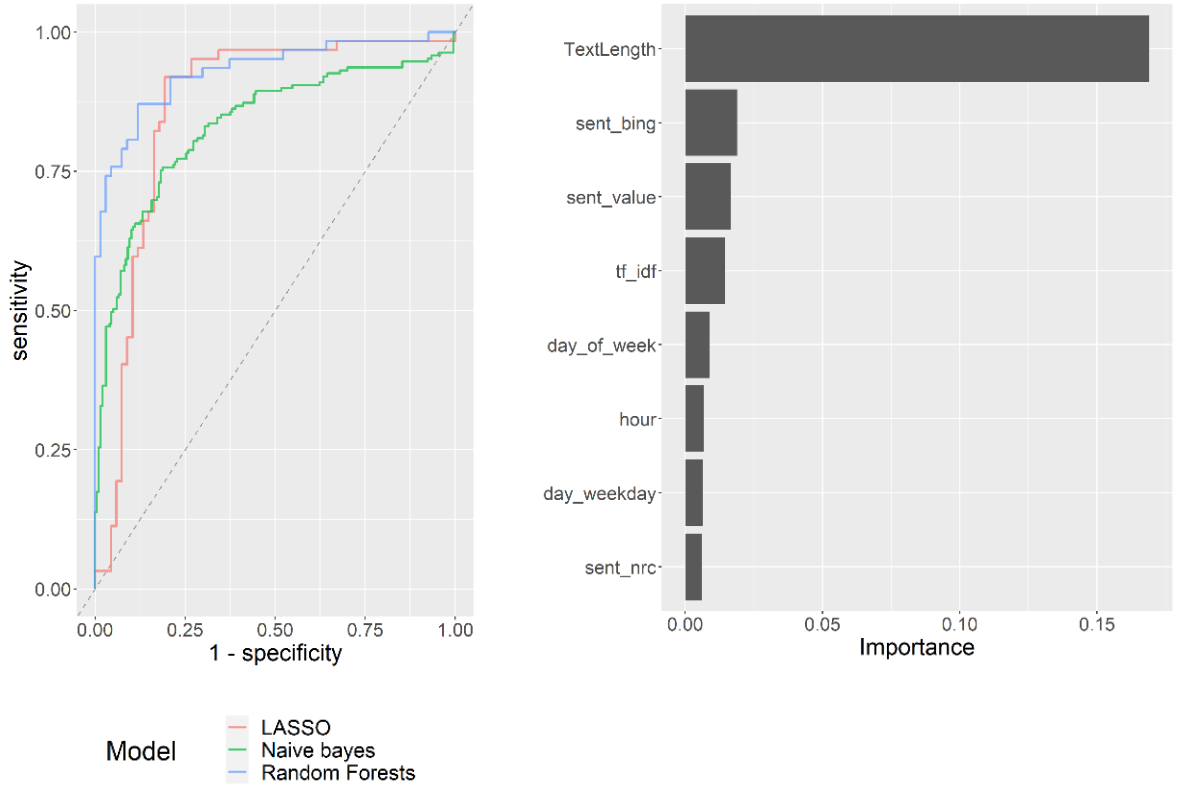


Figure 2. **The left panel** presents the area under the curve (AUC) for three machine learning classification models. As seen here, the Random Forest classification model provides the best performance. **The right panel** illustrates a permutation-based importance test. This examines the level of reliance of Random Forest classification on different model features. Permutation indicates the ability to randomly permute (or shuffling) the order of variables included in a model. Importance represents the difference in baseline model accuracy (before shuffling variables) and performance on the permuted dataset. Higher importance values indicate higher reliance on that variable to maintain model accuracy.

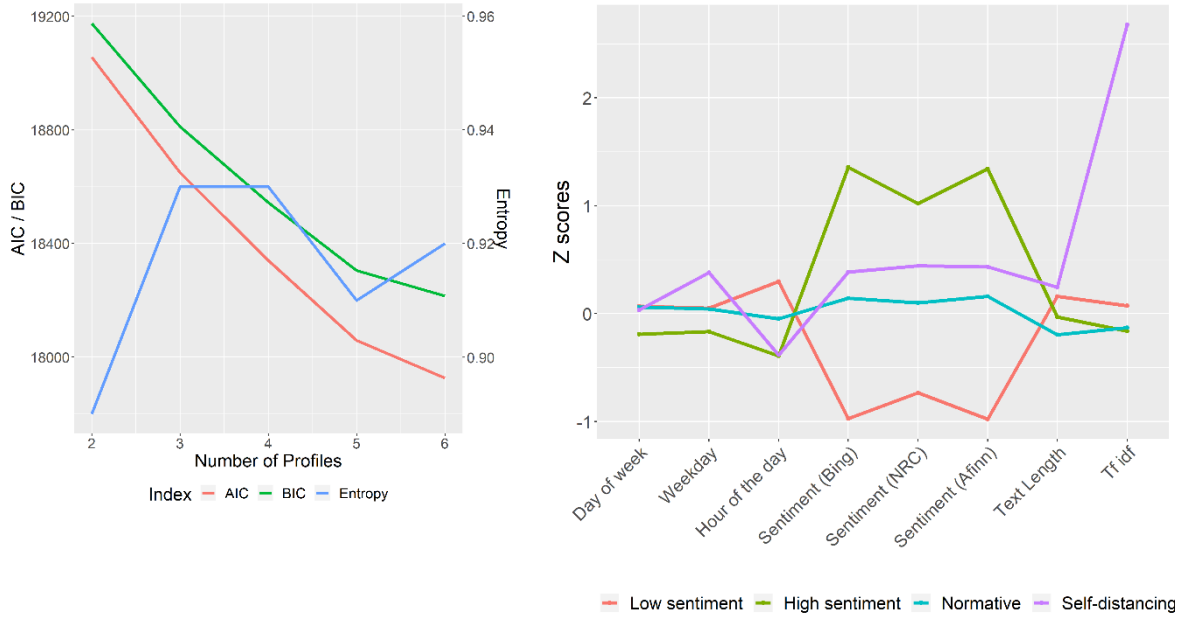


Figure 3. Here the **left plot** helps visualize the identification of optimum number of latent profiles. The left vertical margin shows AIC and BIC values, the right vertical margin shows entropy values, and the horizontal axis shows different number of latent profiles. As seen here, higher number of profiles show lower AIC and BIC; however, entropy is optimized at 4 latent profiles. The **right plot** shows standardized mean scores (vertical axis) of each model indicator (horizontal axis) discriminated by 4 different latent profiles. The low and high sentiment profiles are characterized by sentiment values 1SD below and above the mean, respectively. The self-distancing profile is characterized by a different term frequency.

Supplementary Table 1. Description of Tweets

Variables	Anxious	Control
Number of users	300	305
Number of Tweets	96467	136437
Text length		
Min/Max	0/687	0/834
Mean	91.9	50.1
Text frequency		
Sunday	13588	22007
Monday	12755	19750
Tuesday	14956	21805
Wednesday	13883	17316
Thursday	13629	18754
Friday	13450	15306
Saturday	13112	18934
Weekend	26700	40941
Weekday	68673	92931
Hour (in words)		
0-7	23512	49911
8-15	27432	45099
16-23	45523	41427
Sentiment – Bing		
Positive	45572	73563
Negative	59184	36023
Sentiment – nrc		
Anger	9773	6418
Anticipation	8143	13148
Disgust	7899	6425
Fear	10671	5930
Joy	11162	18067
Negative	18875	11320
Positive	14230	20694
Sadness	10687	6937
Surprise	4661	8194
Trust	8655	13453

Supplementary Table 2. Common positive and negative words by condition

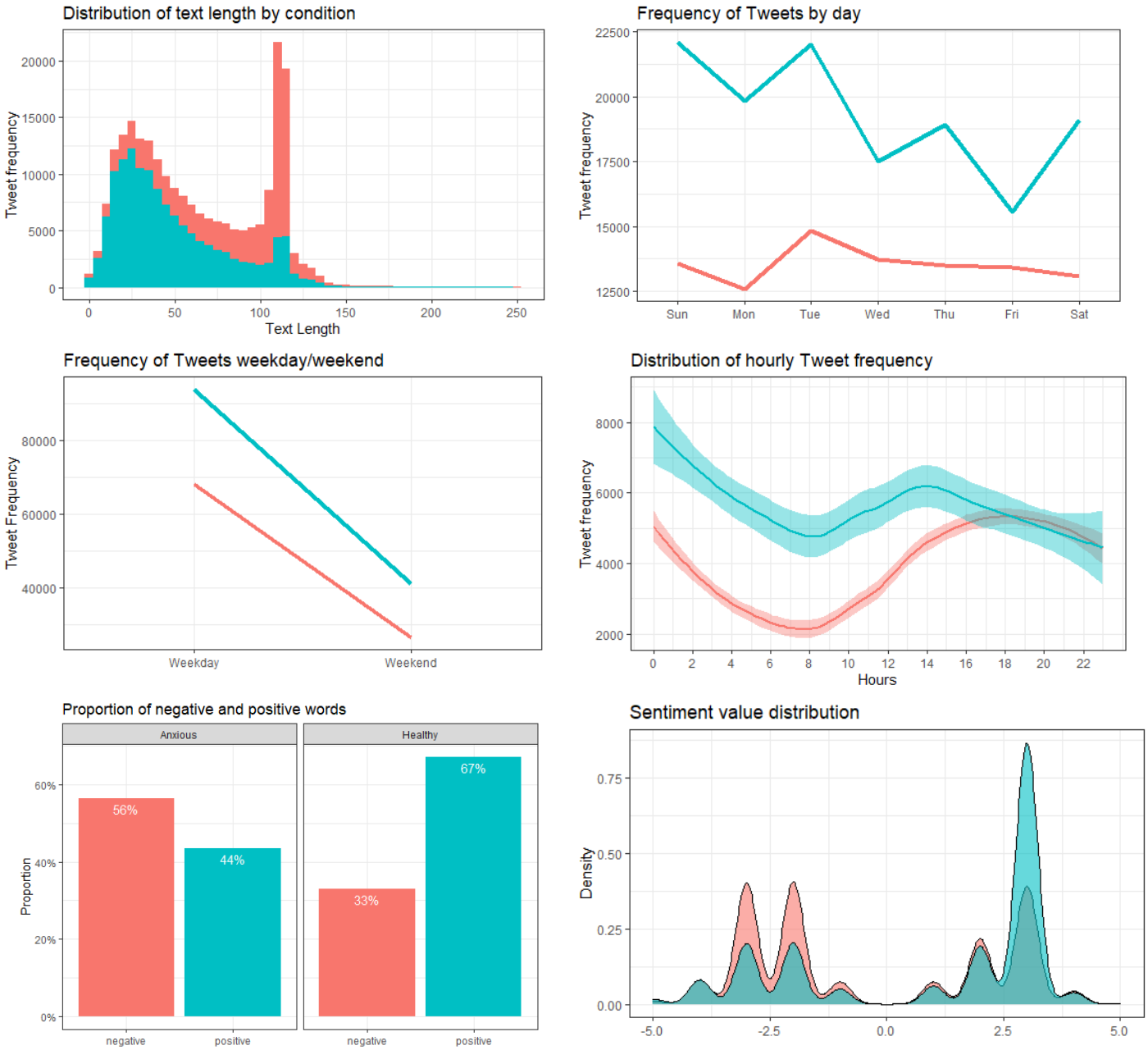
Anxious						Control					
Positive words			Negative words			Positive words			Negative words		
Word	N	Prop	Word	N	Prop	Word	N	Prop	Word	N	Prop
Good	10800	0.2370	Bad	3690	0.0623	Good	24275	0.3300	Bad	3000	0.0833
Love	3642	0.0799	Hate	3140	0.0531	Happy	9728	0.1320	Hate	2640	0.0733
Happy	3156	0.0693	Die	3015	0.0509	Love	5992	0.0815	Hell	1995	0.0554
Pretty	2168	0.0476	Death	1876	0.0317	Pleasant	4385	0.0596	Shit	1932	0.0536
Fun	1137	0.0249	Shit	1764	0.0298	Lovely	3102	0.0422	Damn	1113	0.0309
Honest	1113	0.0244	Hell	1740	0.0294	Pretty	2092	0.0284	Bitch	965	0.0268
Glad	1026	0.0225	Anxiety	1325	0.0294	Beautiful	1524	0.0207	Mad	945	0.0262
Favourite	957	0.0210	Problem	1086	0.0224	Proud	1372	0.0187	Crazy	908	0.0252
Excited	870	0.0191	Evil	880	0.0183	Excited	1230	0.0167	Sick	627	0.0174
Perfect	812	0.0178	Mad	800	0.0149	Fun	1155	0.0157	Pain	606	0.0168
Enjoy	796	0.0175	Crazy	788	0.0135	Perfect	1048	0.0142	Death	574	0.0159
Top	780	0.0171	Damn	780	0.0133	Sweet	900	0.0122	Evil	450	0.0125
Lovely	750	0.0165	Awful	745	0.0132	Top	891	0.0121	Hurt	420	0.0117
Sweet	745	0.0163	Illegal	735	0.0126	Enjoy	868	0.0118	Cry	416	0.0115
Safe	726	0.0159	Worse	627	0.0106	Cute	838	0.0114	Die	393	0.0109
Wonderful	660	0.0145	Murder	618	0.0104	Luck	828	0.0113	Weird	392	0.0109
Proud	624	0.0137	Terrible	615	0.0104	Favourite	780	0.0106	Problem	384	0.0107
Beautiful	582	0.0128	Lost	592	0.0100	Glad	654	0.0089	Kill	381	0.0106
Important	470	0.0103	Fear	564	0.0095	Safe	573	0.0078	Lost	368	0.0102
Luck	464	0.0102	Doubt	524	0.0088	Smile	540	0.0073	Terrible	365	0.0101

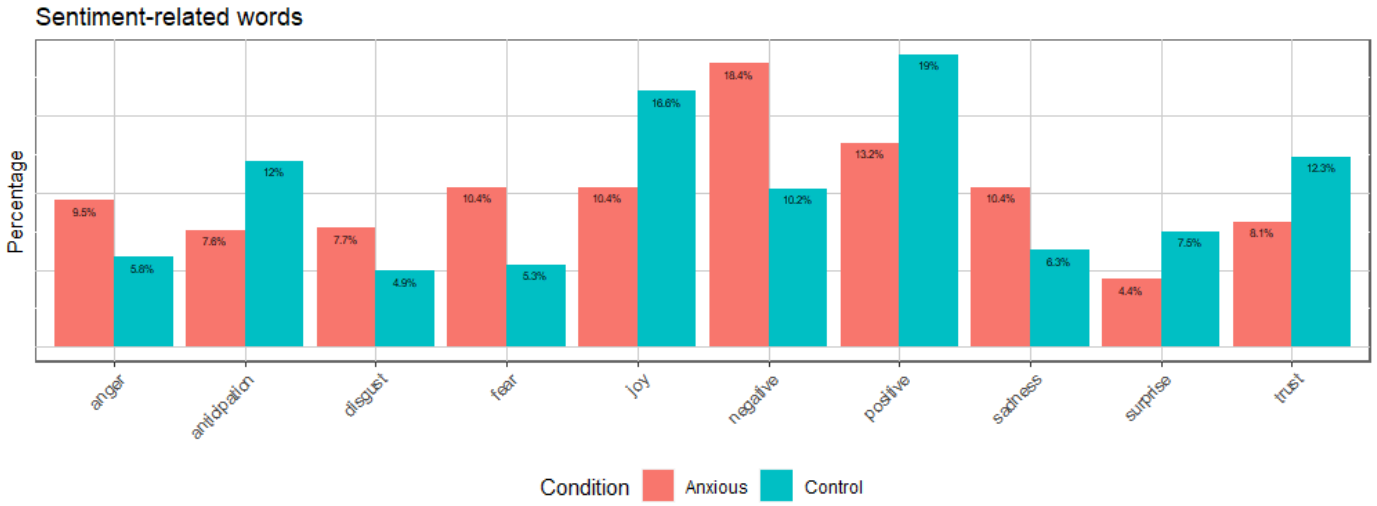
Supplementary Table 3. Selecting optimum number of profiles

Model	Profiles	AIC	BIC	Entropy	N min	BLRT–*p*
1	2	19056	19175	0.89	41%	0.01
1	3	18650	18811	0.93	27%	0.01
1	4	18340	18544	0.93	2%	0.01
1	5	18059	18305	0.91	2%	0.01
1	6	17926	18215	0.92	2%	0.01

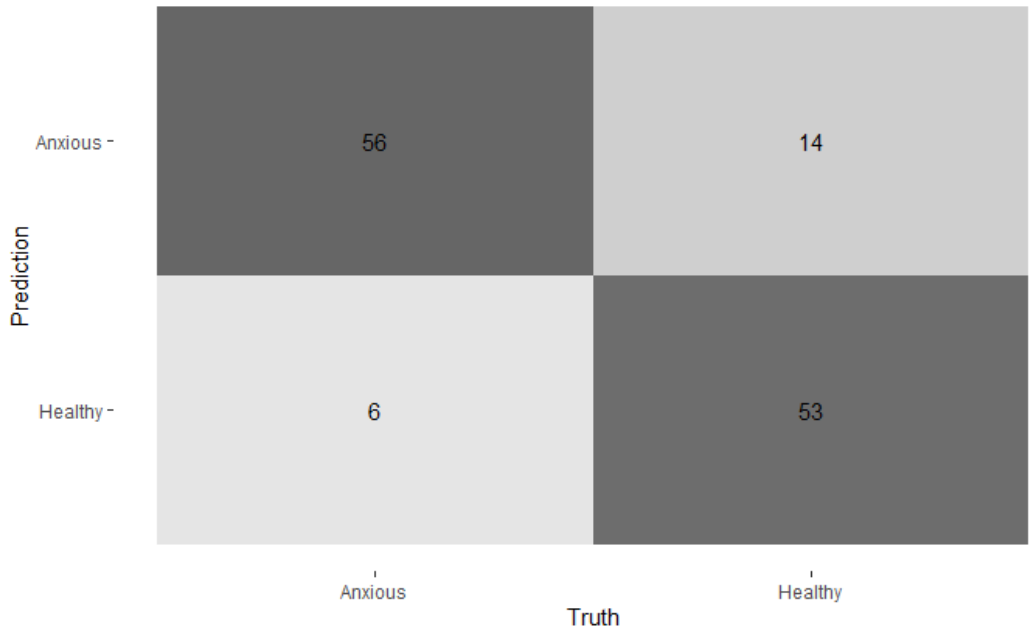
Note: This table presents performance indicators for models with different number of classes. Model 1 is known as Class-invariant diagonal parameterization (CIDP), and it indicates that profile variances are equal and covariances are fixed to zero. AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) are measurements of error with smaller values indicating better fit. Standardized entropy ranges from 0 to 1 and is a measure of heterogeneity across profiles, with higher values indicating clearly distinguishable profiles. N min indicates the percentage of individuals belonging to the smallest latent profile. Bootstrapped likelihood ratio test assesses if adding one more profile to the model represents a significant increment in information.

Supplementary Figure 1. Visualization of user behavior.

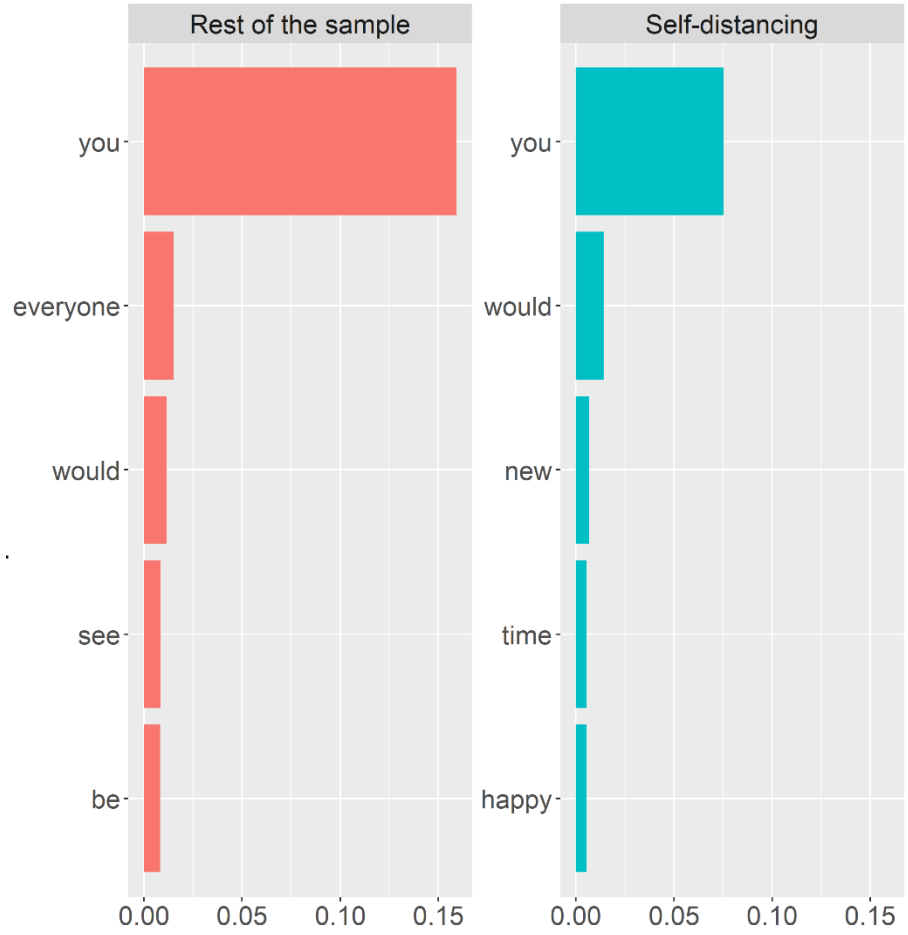




Supplementary Figure 2. Random Forests confusion matrix.



Supplementary Figure 3. Comparison of the five words most frequently used by the 'self-distancing profile and the rest of the sample


Rest of the sample
you
everyone
would
see
be
0.00
0.05
0.10
0.15
Self-distancing
you
would
new
time
happy
0.00
0.05
0.10
0.15