Characterising Mood Instability on Twitter

Anonymous ACL submission

Abstract

The high prevalence of mental health disorders highlights the importance of early detection. Researchers have demonstrated that mood instability is a key characteristic of common mental health conditions like depression and anxiety. In this study, we investigate the distinguishing features, focusing on mood instability, between individuals with and without mental health disorders on Twitter. We utilised an existing dataset, the Twitter Self-reported Temporally-contextual Mental Health Diagnosis. Mood features, including fluctuation, tendency, and episodes, were derived from sentiment analysis over temporal shifts and used in predictive analysis. Our result shows a significant difference between individuals with and without mental health disorders. Particularly, higher mood fluctuation is observed across all mental health disorders. Further, predictive analysis demonstrates a performance of 73% in binary classification and 26% in multi-class.

1 Introduction

Mood instability, manifesting as fluctuating emotional states, is one of the key aspects of mental health. It is reported in 40–60% of individuals with particular mental health disorders (Broome et al., 2015). While mood instability becomes a marker for mental health disorders, traditional methods of assessment can be challenging for early detection. Alternatively, previous research showed significant differences in mood markers, derived from textual features, between individuals with and without mental health disorders in social media (Saha et al., 2017; Saravia et al., 2016). In this study, we aim to investigate the distinguishing features, focusing on mood instability, on Twitter to better recognise mental health disorders.

2 Methods

The research utilised the Twitter Self-reported Temporally-contextual Mental Health Diagnosis (Twitter-STMHD) (Suhavi et al., 2022). The dataset comprises tweets from 26K users self-disclosing a mental health disorder (condition group) and 8K without such disclosures (control group), spanning 2017-2021. It focuses on depression, major-depressive disorder (MDD), postpatrum depression (PPD), post-traumatic stress disorder (PTSD), attention-deficit/hyperactivity disorder (ADHD), obsessive-compulsive disorder (OCD), anxiety disorder, and bipolar disorder.

To quantify mood features, we employed sentiment analysis using Empath (Fast et al., 2016). We define daily mood as the difference between the aggregated mean of positive and negative sentiments per user per day ($\bar{x}^i_+ - \bar{x}^i_-$). We extracted three types of features per user; (1) mood fluctuations (moving standard deviation/MSD): the average standard deviation of $k = 2$ window of days, (2) mood tendencies (mood ratio): the proportion of days with good/bad mood, (3) episodes of mania and depression (mood streak): the frequency of six-days good/bad mood streaks. We compared the distribution in each condition group with the control group using an independent sample t-test.

We further evaluated the features in predictive analysis with 80% hold-out validation, stratified by condition groups. We performed binary classification, combining condition groups into one label ($y = 1$), and multi-class classification, treating each condition group as separate labels. We employed tuned classifiers including *logistic regression* (LR), *naïve bayes* (NB), *support vector machine* (SVM), and *random forest* (RF).

3 Results

Table 1 presents the features analysis result. The result shows high mood fluctuation markers in condition groups as evidenced by higher average MSD. The result is able to capture a pattern of hypomania/mania in individuals with bipolar, as well as

	PPD	MDD	OCD	PTSD	ADHD	Bipolar	anxiety	depression
average MSD	***	***	***	***	***	***	***	***
positive ratio	***	***	***	***	***	***	***	***
negative ratio	***	***	***	***	***	***	***	***
positive streak	-	-	-	-	**	***	***	*
negative streak	-	-	-	*	***	***	***	**

Table 1: Statistical significance of mood features feature (*** < 0.001, ** < 0.01, * < 0.05, . < 0.1)

anxiety and depression groups. Additionally, our result shows a significant difference in both positive and negative tendencies across all condition groups. A higher depressive sign (negative streak) is also observed among PTSD, ADHD, bipolar, anxiety, and general depression groups.

	Binary			Multi-class		
Model	Pr	Rc	F1	Pr	Rc	F1
LR	0.76	0.78	0.71	0.17	0.28	0.18
NB	0.65	0.76	0.67	0.15	0.25	0.17
SVM	0.76	0.78	0.71	0.21	0.29	0.20
RF	0.76	0.79	0.73	0.24	0.28	0.26

Table 2: Classification performance

Next, we tested the mood features in classification models. The binary model yields a performance of 67-73% F1-score (weighted average). Meanwhile, multi-class model performs at 17-26% F1-score. Overall, RF gives the best performance of 73% for binary classification and 26% for multi-class classification as seen in Table 2.

4 Discussion

From the result, we observed that the overall features show significant differences from those in the control group. We also observed that all condition groups show higher mood fluctuations, aligned with the statement from Broome et al. (2015). In addition to that, the condition groups overall can be divided into two; higher mood tendency with and without differences in mood streak. The first one (with) shows characteristics of mood episodes, including bipolar, anxiety, and depression. The second one (without) can be explained by two possibilities: (1) frequent mood-shifting with less dense episodes, and (2) less mood-shifting with longer exposure to mood episodes. As for higher positive tendencies across condition groups, there is a possibility that individuals present positive images of themselves as a coping mechanism. Further sampling may be needed to explore these mood patterns.

The classification model performs well in distinguishing individuals with at least one mental health disorder. However, the multi-class model tends to classify individuals into PPD and bipolar labels as evidenced by higher false positive values for both labels, causing the performance to drop. Further analysis of these labels may give valuable insights into multi-class classification. Overall, the result extends findings in previous research (Saravia et al., 2016), further demonstrating that the features show valuable indicators of mood instability.

References

Matthew R. Broome, Kate E. Saunders, Paul J. Harrison, and Steven Marwaha. 2015. Mood instability: Significance, definition and measurement. *The British Journal of Psychiatry*, 207:283–285.

Ethan Fast, Binbin Chen, and Michael S. Bernstein. 2016. Empath: Understanding topic signals in large-scale text. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, page 4647–4657, New York, NY, USA. Association for Computing Machinery.

Koustuv Saha, Larry Chan, Kaya De Barbaro, Gregory D. Abowd, and Munmun De Choudhury. 2017. Inferring Mood Instability on Social Media by Leveraging Ecological Momentary Assessments. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(3):1–27.

Elvis Saravia, Chun-Hao Chang, Renaud Jollet De Lorenzo, and Yi-Shin Chen. 2016. MIDAS: Mental illness detection and analysis via social media. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 1418–1421, San Francisco, CA, USA. IEEE.

Suhavi, Asmit Kumar Singh, Udit Arora, Somyadeep Shrivastava, Aryaveer Singh, Rajiv Ratn Shah, and Ponnurangam Kumaraguru. 2022. Twitter-STMHD: An Extensive User-Level Database of Multiple Mental Health Disorders. *Proceedings of the International AAAI Conference on Web and Social Media*, 16:1182–1191.