9^{th} International Conference on Computational Social Science IC^2S^2
July 17-20, 2023, Copenhagen, Denmark

Does Papageno Effect Occur on Social Media?

Keywords: social media, mental health, suicidal ideation, causal inference, Papageno effect

Globally, approximately 700,000 people fall victim to suicide each year. The Papageno effect concerns how media can play a positive role in preventing and mitigating suicidal ideation and behaviors [1]. This means that individuals with suicidal ideation are assumed to be positively impacted by seeing how others are coping or have overcome their suicidal thoughts. With the widespread use of social media, individuals often express and share their lived experiences and mental health struggles on these platforms. However, there is a gap in our understanding of the existence and effectiveness of the Papageno effect in social media. In this work, we target the research questions of, *whether the Papageno effect exists on Twitter and how we can quantify psychosocial changes of Twitter users before and after engaging with Twitter posts containing mental health coping stories. We use a causal framework to address these questions.*

To build a Treatment dataset we follow several steps (Figure 1). We collect 13,022 Twitter posts containing at least one term related to suicide attempts and at least one of the terms indicating successful coping between 1 January 2018 and 1 March 2022. Applying a coping story classifier from [2] to annotate the dataset, we find 3,077 coping story posts, of which 709 have replies (Table 1). Replying to coping story posts, indicate that an individual has read and interacted with these stories. We collect 787K Twitter posts from 2,468 individuals who have replied to the 709 coping story posts. These posts constitute our Treatment dataset. We build a Control dataset of individuals who do not reply to coping story posts during the same investigated period. We use keywords, including "life", "job", "music", and "movie" to search for individuals on Twitter. We assign a placebo date from the non-parametric distribution of treatment date in the Treatment dataset to any day the Control individual replies to other Twitter posts to reduce any temporal confounds (Figure 2a). We collect timeline data two weeks before the placebo date and two weeks after of 8,465 individuals to build the *control* dataset.

To examine the psychosocial effects of engaging with coping story posts on social media, we measure three types of psychosocial outcomes drawing from psychiatry and psychology literature: affective, behavioral, and cognitive outcomes [3]. Affect is defined as any experience of feeling or emotion [4]. We use language to infer affective psychosocial wellbeing, using the Linguistic Inquiry and Word Count (LIWC) lexicon [5] and symptomatic mental health expressions classifiers based on transfer learning methodologies [6]. We measure normalized occurrences of words in affective categories and the aggregated proportion of expressing mental health concerns per individual. Behavioral outcomes are defined as an individual's overt actions, behavioral intentions, and verbal statements regarding behavior [3]. We use activity, interactivity, and topic diversity metrics. Activity is the average number of Twitter posts per day, interactivity is the proportion of replies to original Twitter posts, and topic diversity is calculated using a language model on posted texts by measuring the average cosine distance from the centroid of the corresponding corpus. Cognitive outcomes relate to beliefs, knowledge structures, perceptual responses, and thoughts [3]. We use readability, complexity and repeatability, and psycholinguistic keywords to measure cognitive behaviors. Readability is measured using the Coleman-Liau Index [7], while repeatability and complexity are calculated based on the normalized count of non-unique words and the average number of words per sentence. We analyze psycholinguistic keywords using the LIWC lexicon, focusing on five aggregated categories.

Our aim is to measure the psychosocial outcomes of engaging with a coping story post using propensity score matching to match Treatment and Control individuals with similar pre-Treatment behavior. We build several covariates from the Treatment and Control datasets to control for similar pre-Treatment behavior on social media, including social media features, word usage distribution, psycholinguistic features, and average usage of posts related to symptomatic mental health expressions. Matching is employed to pair Treatment individuals and Control individuals with similar covariates. After matching, there are 1,245 Treatment and 1,087 Control individuals, and we measure the balance of the covariates using the standardized mean differences (SMD) between the two groups. In our study, the maximum SMD is 0.19, and the mean SMD is 0.06, which means that the two groups are very similar [6].

We present the shifts for each psychosocial outcome across the matched Treatment and Control individuals in the corresponding datasets. We calculate the effect size (Cohen's d), and measure statistical significance in differences using an independent sample t-test. In table 2, we observe that engaging with coping story posts on Twitter is linked to lower stress and depression, and higher usage of affect words, expressive writing, diversity, and interactivity. Our findings suggest that engaging with posts featuring coping stories can have positive impacts on psychosocial wellbeing and provide evidence for the Papageno effect on social media. Our work proposes a methodology for measuring these outcomes, highlighting the potential role of social media in preventing suicide. This has practical implications for developing strategies to support vulnerable members in online communities and for designing social media platforms to encourage positive behavior.

References

[1] Niederkrotenthaler T, Voracek M, Herberth A, Till B, Strauss M, Etzersdorfer E, et al. Role of media reports in completed and prevented suicide: Werther v. Papageno effects. The British Journal of Psychiatry. 2010;197(3):234-43.

[2] Metzler H, Baginski H, Niederkrotenthaler T, Garcia D. Detecting Potentially Harmful and Protective Suicide-related Content on Twitter: A Machine Learning Approach. arXiv preprint arXiv:211204796. 2021.

[3] Breckler SJ. Empirical validation of affect, behavior, and cognition as distinct components of attitude. Journal of personality and social psychology. 1984;47(6):1191.

[4] VandenBos GR. APA dictionary of psychology. American Psychological Association; 2007.

[5] Pennebaker JW, Chung CK. Expressive writing, emotional upheavals, and health. Handbook of health psychology. 2007:263-84.

[6] Saha K, Sugar B, Torous J, Abrahao B, Kıcıman E, De Choudhury M. A social media study on the effects of psychiatric medication use. In: Proceedings of the International AAAI Conference on Web and Social Media. vol. 13; 2019. p. 440-51.

[7] Ernala SK, Rizvi AF, Birnbaum ML, Kane JM, De Choudhury M. Linguistic markers indicating therapeutic outcomes of social media disclosures of schizophrenia. PACM Human-Computer Interaction. 2017;(CSCW).

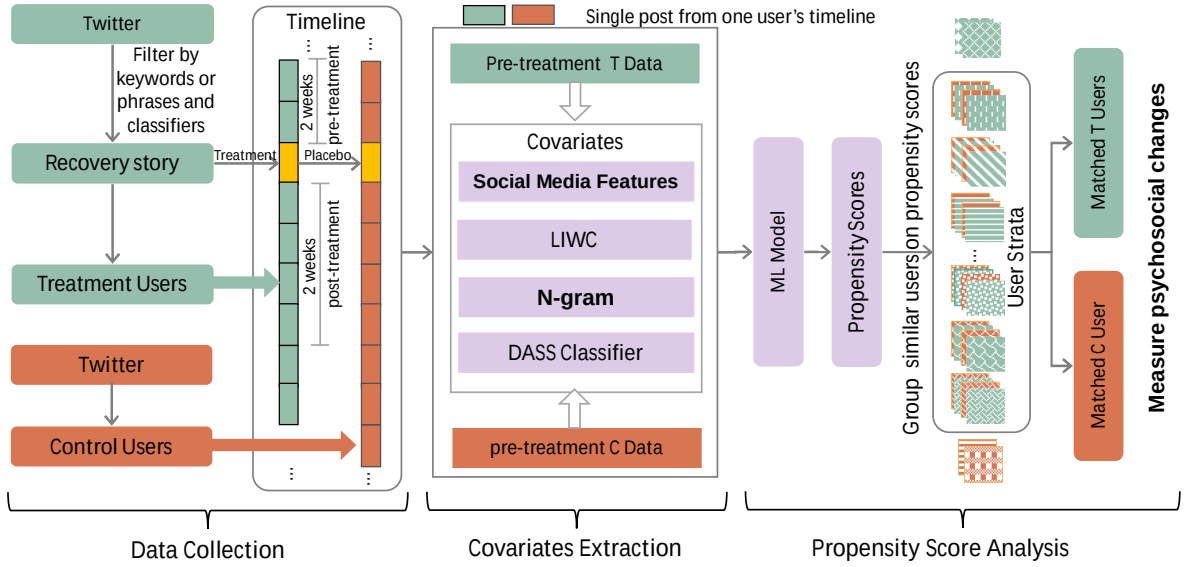
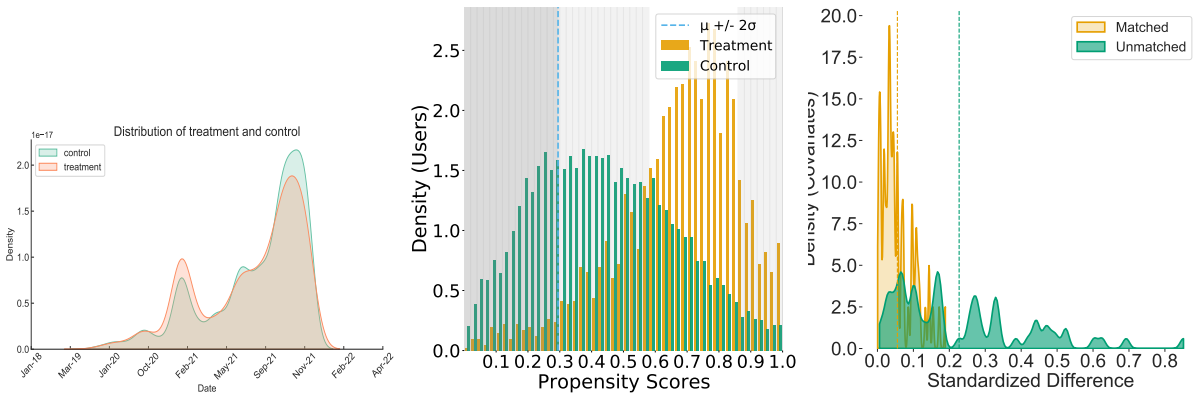


Figure 1: Schematic diagram of propensity score matching between Treatment individuals and Control individuals.



(a) (b) (c)

Figure 2: (a) Treatment and Control (placebo) dates distribution, (b) Propensity score distribution (shaded region are the dropped strata). (c) Quality of matching.

Table 1: Paraphrased example Twitter posts labeled with coping story or non-coping story and Twitter posts responding to coping story posts. We manually verified the accuracy of the classifier on a sample of coping story posts, which indicated an 86.7% accuracy rate.

Coping Story Posts
"I'm posting this because I've had suicide ideas passively for a long time. I finally realized I was suicidal three years ago. I believed that the desire to be better off dead was common. It is NOT the norm. If you have such ideas, you should seek professional assistance."
"When I was a patient in the psychiatric hospital, they had to remove my shoelaces to prevent me from self-injury. Today marks one month without suicide ideation. My life has improved after receiving therapy from all of my physicians. Cheers to the continuation of living in the present!"
Non-Coping Story Posts
"Even more terrible than my thoughts of death are my suicide ideas. People told me when I was 10 that it would get better, but it hasn't, and I want to die yet nothing works. It's so unfair that no matter how many times I try, I always fail. I'm sorry if this is frustrating; I just feel so alone."
"But I wanted to kill myself again this weekend. I've never been happier. But every day is so full of grief for the body I don't have and will never be capable of having."
Twitter Posts Responding to Coping Story Posts
"Dear friend, I'm happy that things didn't turn out the way you had hoped. I wouldn't want the past few of years to have been any other way because they have been such an adventure. To where we all go in the future is something I am looking forward to see."
"I'm glad to hear that you're doing well. It's good to know that you have support from close friends and family because I am aware of how challenging it may be to handle some circumstances."

Table 2: Summary of psychosocial differences across all the outcomes between Treatment and Control individuals. We report mean psychosocial outcomes across all matched individuals, effect size (Cohen's *d*), independent sample t-statistic. The p-values from LIWC categories are adjusted using non-negative two stage FDR correction ($*p < 0.05, **p < 0.01, ***p < 0.001$).

Categories	Tr.	Ct.	RTE	d	t-test
Affective Outcomes					
LIWC: Affect	0.09	0.08	1.05	0.20	2.19*
Anxiety	0.05	0.04	1.08	0.17	-0.33
Depression	0.15	0.17	0.93	0.28	-2.84**
Stress	0.35	0.38	0.94	0.28	-3.96***
Suicidal Ideation	0.07	0.06	1.04	0.15	-0.41
Behavioral Outcomes					
Activity	4.33	4.19	1.10	0.16	0.40
Interactivity	8.89	2.78	3.34	0.34	4.17**
Topics Diversity	0.37	0.36	1.02	0.25	2.51*
Cognitive Outcomes					
Readability	12.33	11.52	0.95	0.18	-2.16*
Complexity	9.26	9.52	0.99	0.19	-1.42
Repeatability	0.51	0.45	1.11	0.30	5.09***
LIWC: Cognition & Perception	0.27	0.27	1.02	0.17	1.12
LIWC: Social Context	0.18	0.17	1.01	0.08	0.55
LIWC: Lexical Density & Awareness	0.60	0.61	0.99	0.15	-1.23
LIWC: Interpersonal Focus	0.12	0.11	1.08	0.26	2.77*
LIWC: Temporal Reference	0.10	0.10	1.02	0.17	1.45