Mental Health Coping Stories on Social Media: A Causal-Inference Study of Papageno Effect

Yunhao Yuan

School of Science

Thesis submitted for examination for the degree of Master of Science in Technology.
Espoo 03.11.2022

Supervisor

Dr. Barbara Keller

Advisor

Dr. Talayeh Aledavood

Aalto University
School of Science

Aalto University
School of Science

Aalto University, P.O. BOX 11000, 00076 AALTO
www.aalto.fi
Abstract of the master's thesis

Author Yunhao Yuan

Title Mental Health Coping Stories on Social Media: A Causal-Inference Study of Papageno Effect

Degree programme Master's Programme in Computer, Communication and Information Sciences

Major Computer Science **Code of major** SCI3042

Supervisor Dr. Barbara Keller

Advisor Dr. Talayeh Aledavood

Date 03.11.2022 **Number of pages** 51 **Language** English

Abstract

The Papageno effect concerns how media can play a positive role in preventing and mitigating suicidal ideation and behaviors. With the increasing ubiquity and widespread use of social media, individuals often express and share lived experiences and struggles with mental health on these platforms. However, there is a gap in our understanding about the existence and effectiveness of the Papageno effect in social media, which are studied in this thesis. In particular, this work adopts a causal-inference framework to examine the impact of exposure to mental health coping stories on individuals on Twitter. A Twitter dataset with ∼2M posts by ∼10K individuals is obtained. This work considers engaging with coping stories as the Treatment intervention, and adopts a stratified propensity score approach to find matched cohorts of Treatment and Control individuals. This work measures the psychosocial shifts in affective, behavioral, and cognitive outcomes in longitudinal Twitter data before and after engaging with the coping stories. The findings of this study reveal that, engaging with coping stories leads to decreased stress and depression, and improved expressive writing, the diversity of expressed topics in posts, and interactivity. This thesis discusses the practical and platform design implications in supporting mental wellbeing.

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

Preface

This work begins with the motivation of my interest in suicide study. During this project, one of my friends chose to leave this world because of the bipolar disorder he had suffered from for a long time. With all kinds of emotions and feelings, I thought the best way to memorize him was to finish this project about suicide prevention.

Many people deserve my heartfelt appreciation for supporting me in various ways during this project. My deepest thanks go to my thesis instructors, Dr. Talayeh Aledavood, Dr. Koustuv Saha, and Dr. Barbara Keller. They have been with me with study guidance and welling-being support. Besides, I would like to thank Dr. Erkki Isometsä, who provides valuable insights from the perspective of psychiatry. My thanks go to Dr. Hannah Metzler, for her kindness in giving technical support and encouragement. In the end, I would like to thank my friends and family, who have provided me with support, patience, and love.

Otaniemi, 03.11.2022

Yunhao Yuan

Contents

Abbreviations

WHO	World Health Organization
EU	European Union
SUTVA	Stable Unit Treatment Value Assumption
ATR	Average Treatment Effect
RTE	Relative Treatment Effect
ITE	Individual Treatment Effect
LIWC	Linguistic Inquiry and Word Count
SVM	Support Vector Machine
TF-IDF	Frequency-Inverse Document Frequency
CNN	Convolutional Neural Network
LSTM	Long Short-Term Memory
DNN	Deep Neural Net
API	Application Programming Interface
KS test	Kolmogorov-Smirnov test
API	Application Programming Interface
CLI	Coleman-Liau Index
SMD	Standardized Mean Differences
LG	Logistic Regression

1 Introduction

1.1 Motivation

According to a report from the World Health Organization (WHO), (WHO 2019), globally, approximately 700,000 people fall victim to suicide annually. Suicide is the third leading cause of death among those aged between 15 to 19, and it is the second reason for deaths in 15- to 29-year-olds. A report from Eurostat[1] observes that 1.1% of the 5.2 million fatalities recorded in the European Union (EU) are the result of suicide in 2015. Almost eight out of ten suicides involve males, and around 31% involve individuals aged 45 to 60. The suicide rate at the EU level is 11 per 100 000 persons. Lithuania has the highest suicide rate in Europe, with 30 suicide cases per population of 100,000 annually. Non-fatal suicidal behaviors are more common than suicides. For each suicide, there are around 20-30 suicide attempts. Based on the data from 108 705 participants in the WHO world mental health survey (Borges et al. 2010), the average 12-month prevalence of suicidal thoughts is 2% in high-income countries and 2.1% in developing countries. In contrast, the prevalence of suicide attempts is 0.3% and 0.4%, respectively. Suicide attempts and particularly committed suicides cause severe and tragic consequences among relatives and friends of the victims, as well as significant economic problems for society. Consequently, suicide has become a crucial global public health problem, and WHO has called for urgent action to reduce suicide mortality.

Suicide is a complex issue interacting with various risk factors throughout the lifetime. Previous research divides suicidal factors into personality and individual differences, cognitive factors, social aspects, and negative life events (O'Connor and Nock 2014). The existence of past mental conditions is the most well-researched risk factor for suicidal behavior. More than ninety percent of suicide victims have mental health disorders prior to their attempts (Cavanagh et al. 2003). Another research finds rates of suicidal attempts are higher in young individuals, women, the unmarried, and socially disadvantaged people (Hawton and van Heeringen 2009). Physical illnesses also contribute to suicidal behavior. Significant associations exist between the existence and accumulation of physical ailments (such as heart disease, chronic pain, and respiratory problems) and eventual suicidal behavior (Scott et al. 2010).

While suicide is a combined outcome of multiple, interrelated factors, ranging from mental health issues to social factors, media can play an important role either in a harmful or beneficial direction. The harmful effect of media, dubbed the "Werther effect" (Phillips 1974), refers to the association between the media describing suicides by celebrities and a series of related copycat suicides in the general population. A considerable amount of literature has studied and re-confirmed the Werther effect across different periods, and geographic regions (Fahey, Matsubayashi, and Ueda 2018; Stack 1987; Wasserman 1984). Recent work estimates that the suicide rate increases 8-18% in 2 months after media reporting about celebrity suicides (Niederkrotenthaler et

[1]Suicide rate in the EU in 2015: https://ec.europa.eu/eurostat/web/products-eurostat-news/

al. 2020). By understanding the Werther effect, mental health and suicide prevention organizations around the world, including the World Health Organization, produces standards for responsible suicide reporting by the media, with a particular focus on news and media (Organization and others 2014).

Although extensive research has been carried out on the harmful effect of media, there is much less information about the beneficial effect of media. A study from 2010 (Niederkrotenthaler et al. 2010) explores, for the first time, the possible protective effect of media reporting suicide, referred to as the "Papageno effect". This study analyzes the content of 490 newspaper articles on suicide in Austria for six months. It finds a decrease in suicides after reporting, which portrays ways of overcoming suicidal ideation without narrating suicidal behaviors. This work provides important insights into the potential benefits of media reporting suicide in stories of hope and recovery. Following this work, other studies provide evidence of the Papageno effect from fictional films (Till et al. 2015), suicide-educational websites (Till et al. 2017), and newspaper articles (Arendt, Till, and Niederkrotenthaler 2016). Given the prevalence and importance of social media, understanding more about the Papageno effect on social media can play a crucial role in decreasing suicide rates.

Studies of the Papageno effect commonly rely on self-reports, surveys, and publicly reported suicide statistics and only cover a small, selected group of people. People with suicidal ideation can face negative attitudes and stigmatization, which prevents many of them from seeking help (Reynders et al. 2015). Additionally, the sensitive nature of suicide makes it challenging to collect data at scale on people who have suicidal behaviors and conduct continuous follow-up studies over a long time.

The emergence of social media platforms, such as Twitter, Reddit, and TikTok, provides venues for people to express themselves and connect with others. Various studies utilize data from different social media platforms to explore psychological and health issues using data from various fields such as drug misuse (Garg et al. 2021), minority stress (Yuan et al. 2022), and mental health (Coppersmith et al. 2018). Social media platforms provide timely and relevant information on tracking risk attributes longitudinally. A recent study (Saha et al. 2021b) shows that people tend to disclose information on social media, which is complementary to what they disclose otherwise. The anonymous features of social media may reduce the biases found in research based on surveys and self-reported data. As a result, social media data provide an unparalleled chance to research suicide ideation in the general public and how it changes over time. Utilizing this opportunity is particularly crucial in the context of the present pandemic, when it is well-recognized that mental health difficulties have greatly increased.

1.2 Problem Definition

This thesis uses public data from Twitter, a popular social media site with 330 million daily active individuals worldwide in 2019 (Yeasmin et al. 2022). This work analyzes longitudinal posts from Twitter users who reply to Twitter posts containing stories about coping with suicidal ideation. The psychosocial changes are examined in affective, behavioral, and cognitive outcomes related to suicidal ideation. Specifically,

this work targets the research questions of, *whether the Papageno effect can be observed on social media and how to quantify psychosocial changes of Twitter users before and after engaging in Twitter posts containing mental health coping stories.*

To achieve the research goals, this project collects 13,022 Twitter posts containing keywords, which might indicate coping stories. A machine learning classifier is utilized from a previous study (Metzler et al. 2021) to annotate the dataset automatically. In total, 3,077 Twitter posts are labeled as a coping story, which might result in the Papageno effect. Among them, the author manually verifies the accuracy of the classifier on a sample coping story dataset. This project collects data from two populations on Twitter: 787K posts from 2,468 individuals who reply to Twitter posts containing coping stories and 1.4M posts from 8,465 individuals in a control group. After applying stratified propensity score matching, psychosocial outcomes are aggregated as affective, behavioral, and cognitive outcomes and identify these with highly significant effects. The results show that engaging in coping story posts on Twitter is linked to lower stress and depression, and higher expressive writing, the diversity of expressed topics in posts, and interactivity.

This is the first study that utilizes social media data to quantify the psychosocial shifts of coping stories on social media. The finds have practical and platform design implications in supporting mental wellbeing and preventing potential suicide attempts.

1.3 Outline of Thesis

The rest of this thesis is organized as follows: Section 2 begins with basic concepts of causal inference and then goes on to a review of the related works. In Section 3, the process of compiling datasets is presented in detail. Section 4 is about the study design and methodologies used for this study. Section 5 presents the findings and Section 6 explains how the methods and findings can provide insights into reporting suicide by utilizing the Papageno effect and the existing limitations of this study. Finally, Section 7 concludes the thesis.

1.4 Privacy, Ethics, and Disclosure

All social media data in this work are publicly available. The author of these data was not contacted or interacted with as part of this study. Nevertheless, this work commits to protecting the privacy of individuals by implementing several methodologies. This work anonymizes the data by removing all personally identifiable information and paraphrasing quotes.

2 Background and Related Work

The basic definitions and conceptions of causal inference are shown in Section 2.1. Section 2.2 introduces research about media effects on suicidal ideation. Section 2.3 briefly discusses the associations between mental health and linguistic cues associated with suicide. Section 2.4 represents the prevailing research on suicide using social media, and finally, Section 2.5 covers the current approach to classifying suicidal ideation.

2.1 Casual Inference

In the field of clinical research, randomized controlled trials are widely used to estimate the effectiveness of a new intervention or treatment. However, in some situations, random assignment is not feasible or impossible for different reasons, such as time constraints, practical and ethical constraints, and limited generalizability (Staffa and Zurakowski 2018). One practical alternative approach is called the propensity score matching framework, which estimates the causal effect of a treatment or an intervention (Ali et al. 2019). The propensity score approach utilizes a wide range of covariates to match individuals in a treatment group with individuals with comparable characteristics but in the absence of treatment (Figure 1). This section introduces the preliminaries of causal inference and presents how to calculate propensity score and different propensity score methods.

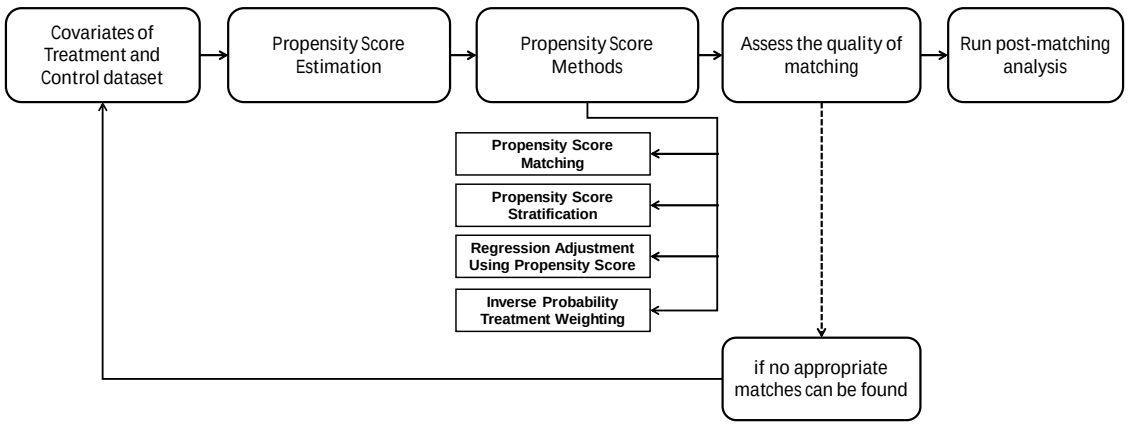


Figure 1: The implementation steps of propensity score framework.

2.1.1 Preliminaries

In this section, the basic notions and concepts in propensity score matching are introduced.

Definition 1. *Unit. A unit is a subject, which a treatment or an intervention will operate.*

A unit is indicated by i, where $i \in \{1, 2, ..N\}$ with the number N subjects.

Definition 2. *Covariates. Covariates are observed characteristics or attributes (except actual treatment) of experiment units.*

For unit i, the observed covariates are indicated as vector x_i.

Definition 3. *Treatment. Treatment is an indicator of treatment or intervention intake for a unit.*

For binary treatment study, the treatment is defined as

$$T_i = \begin{cases} 1 & \text{if unit } i \text{ received the treatment} \\ 0 & \text{otherwise.} \end{cases} . \tag{1}$$

Definition 4. *Potential outcome. A potential outcome is the outcome if a unit i receives a treatment.*

For every unit i, a pair of potential outcomes can be indicated as: (1) Y_{1i}, the potential outcome for unit i under the treatment condition; (2) Y_{0i}, the potential outcome for unit i under the control condition.

Definition 5. *Observed outcome. The observed outcome the outcome under the actual control or treatment condition.*

For every unit i, only one potential outcome can be observed, donated as

$$Y_i = Y_{1i}T_i + Y_{0i}\left(1 - T_i\right). \tag{2}$$

Definition 6. *Counterfactual outcome. The unobserved outcome under the actual control or treatment condition is the counterfactual outcome.*

For unit i in treatment group, when Y_{1i} is observed outcome, Y_{0i} is counterfactual outcome. unit i in control group, when Y_{0i} is observed outcome, Y_{1i} is counterfactual outcome. unit i in control group.

The definition and notions can calculate the treatment effect at different levels, including individual, population, and treatment levels (Yao et al. 2021).

Definition 7. *Individual treatment effect. For unit i, the individual treatment effect is defined as the difference between its two potential outcomes:*

$$ITE_i = Y_{1i} - Y_{0i} \tag{3}$$

The treatment effect over population can be obtained by averaging the ITE.

Definition 8. *Average treatment effect. The average treatment effect is defined as*

$$ATE = E[Y_{1i} - Y_{0i}] \tag{4}$$

The average can be restricted to treated units (ATT: Average Treatment effect on the Treated).

Definition 9. *Average treatment effect on the treated. The Average treatment effect on the condition of treatment is defined as*

$$ATT = E[Y_{1i} - Y_{0i}|T_i = 1] \tag{5}$$

There are other ways to calculate to treatment effect. For example, the treatment effect on individuals can be calculated as an average ratio, defined as relative treatment effect (RTE)

$$RTE = E[Y_{1i}/Y_{0i}] \tag{6}$$

2.1.2 Propensity Score Estimation

When analyzing non-randomized, observational data, propensity score approaches are used to minimize the risk of confounding and lower the bias in calculating treatment effects. One assumption needed to be satisfied before applying propensity score methods is the stable unit treatment value assumption (SUTVA) (Rubin 2005). It requires the treatment effect for one unit should be unrelated to the treatment status of another. In this section, several propensity score methods are discussed under this assumption.

Definition 10. *Propensity score. The propensity score refers to the likelihood that a unit would receive the treatment based on observed covariates.*

$$e(x) = P(t = 1|X = x) \tag{7}$$

The propensity score indicates the probability that a unit is in the treatment group or control group based on the characteristics of the units (covariates).

In practice, logistic regression is the most common method to calculate propensity scores, in which treatment outcomes are regressed on pre-treatment characteristics (covariates). It predicts the likelihood, ranging from 0 to 1, that a unit is exposed to treatment. There are various benefits to employing logistic regression, such as it is simple to implement and it is a well-known and well-understood statistical technique for researchers to interpret (Setoguchi et al. 2008). The definition of using logistic regression to estimate propensity score is donated as:

$$\ln \frac{e\left(x_i\right)}{1 - e\left(x_i\right)} = \ln \frac{\mathrm{P}\left(T_i = 1 \mid x_i\right)}{1 - \mathrm{P}\left(T_i = 1 \mid x_i\right)} = a + \beta^{\top} x_i, \tag{8}$$

where

- $e\left(x_i\right) = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + \ldots + b_i x_i$
- b_0 is the intercept,
- b_i is the regression coefficient,
- x_i is the observed covariates.

2.1.3 Propensity Score Methods

After estimating the propensity, there are different approaches that researchers might use in their design or analysis. There are four general methods to adjust estimated propensity scores, including propensity score matching, propensity score stratification, regression adjustment using propensity score, and inverse probability treatment weighting (Ali et al. 2019). It can be used individually or mixed used.

Propensity Score Matching. The most common propensity score method is propensity score matching, which divides the study units into two groups based on similar or identical propensity scores (Figure 2). One group is exposed to the treatment, and the other one is not (Rozé et al. 2015). The matching can be implemented in different ways, such as one-to-one, one-to-many, and matching with or without replacements (Hansen 2004). After matching, the covariates balance between the treatment and control groups can be estimated by absolute standardized difference. Propensity score matching is mainly used to measure ATT, not ATE, as the remaining untreated people who were not matched with the nearest treated patients are often removed from the study (Schafer and Kang 2008). By directly comparing the examined outcomes across units in the matched treatment group and control, propensity score-matched sample analysis may simulate the design of a randomized controlled trial.

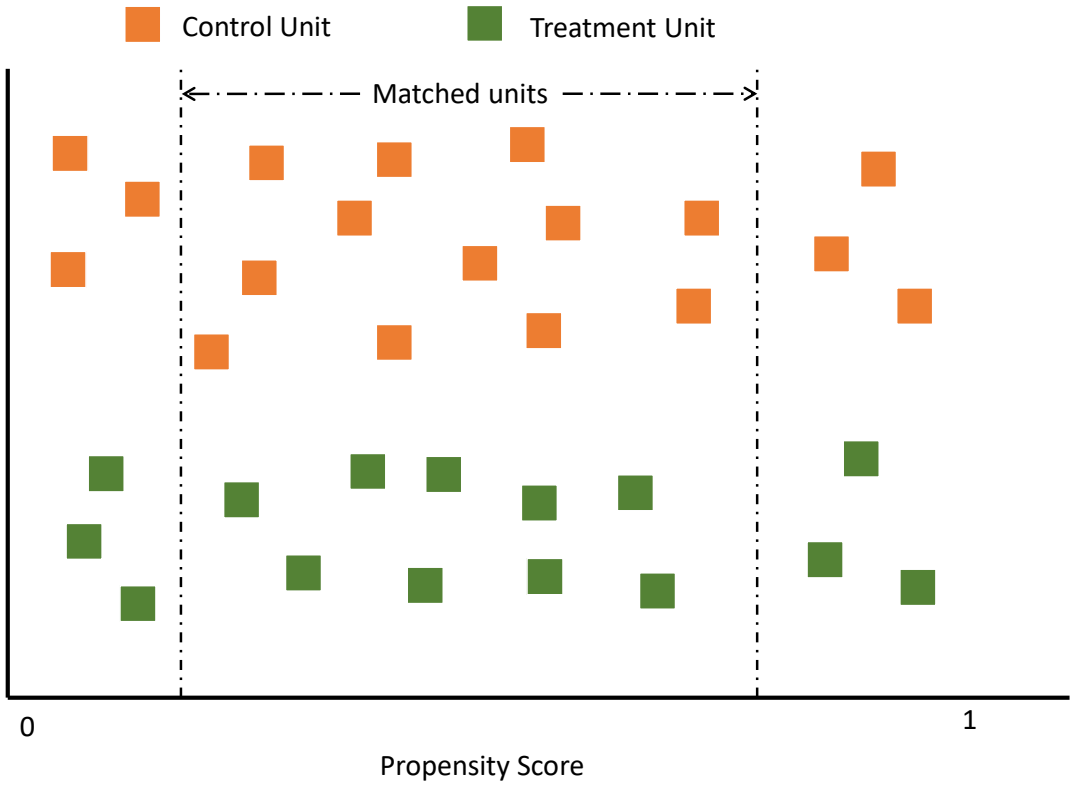


Figure 2: An illustration of propensity score matching.

Propensity Score Stratification. The second approach is based on the stratification of the propensity score (Rosenbaum and Rubin 1983), also named propensity score subclassification. This method is based on the idea that people in each stratum are more similar to one another than people in general. Therefore the outcomes in the same stratum may be compared directly. With the propensity scores, units in treatment and control groups are divided into different subgroups or strata. Within each stratum, treatment and control units are controlled to have a similar distribution across all measured covariates (Rosenbaum and Rubin 1984). By increasing the number of strata, the likelihood of bias will decreases. To calculate an overall estimate of the link between treatment and measured outcomes, using the number of units in each stratum as weight yields the ATE. ATT is calculated by using the ratio of units in the treatment group in each stratum.

Regression Adjustment Using Propensity Score. Regression adjustment based propensity score utilizes a separate multi-variable model, in which the propensity score is the model's input and the measured outcome is the output. Controlling for the likelihood of receiving the treatment enables the researcher to measure the outcome linked to the treatment of interest while lowering confounding. Despite the easy implementation of the regression adjustment method, it is considered as a poor way to use the propensity score as it mixes the study design and data analysis steps (Rosenbaum and Rubin 1983).

Inverse Probability Treatment Weighting. The last method is inverse probability of treatment weighting using the propensity score, also called a "pseudo-population" (Schaffer et al. 2015). The usage of inverse probability of treatment weighting is comparable to the use of survey sampling weights, which are used to ensure that survey samples are representative of certain populations. (Morgan and Todd 2008). For the purpose of creating a sample in which the distribution of possible confounding variables is independent of exposure, statistical weights for each person are calculated using propensity scores. Weights can be defined as

$$w_i = \frac{T_i}{e\left(x_i\right)} + \frac{1 - T_i}{1 - e\left(x_i\right)} \tag{9}$$

It allows an unbiased measurement of the relationship between treatment and outcome (Robins, Hernan, and Brumback 2000). This method is mainly used to ATE across the population.

2.2 Media Effects on Suicidal Ideation

The question of how suicide reporting in the media influences subsequent suicides has received considerable attention. Table 1 summarizes findings on selected digital mood related studies. For quite some time, studies have focused on the negative impact of media portrayals on suicide and found a positive correlation between media coverage of suicidal behavior and suicidality (Domaradzki 2021). Ladwig et al. finds that after the famous German national goalie Robert Enke committed suicide on a train, the overall suicide rate rose by 117.2% (Ladwig et al. 2012). Fu, Yip, and others demonstrate that celebrity suicides represent a separate risk that may result

in suicidal ideation in anyone, not only those who are depressed or going through a difficult time in life (Fu, Yip, and others 2009). Another study contends that non-celebrity suicides might also trigger the imitation effect on those from similar backgrounds or socioeconomic standing (Niederkrotenthaler et al. 2009).

Scholars have long debated the possible preventive impact of media reporting on suicide. The first serious discussions and analyses of the preventive effect of media emerged during the 1960s, Motto first shows that a newspaper blackout, a reduction of the number of reporting, or a change in the quality of reporting style might lower suicide cases. They find the decline in young females who committed suicide after a 286-day newspaper blackout in Detroit (Motto 1970). Similarly, another study (Blumenthal and Bergner 1973) shows a decrease in suicide rates after newspaper strikes prevented news reports from being published. Holding find no increase in suicide attempts after a TV show about a suicidal person whom a suicide prevention agency helps in Edinburgh. The research also discovers a sharp rise in new client referrals to the Edinburgh branch, which suggests that the TV show raises public awareness of the suicide prevention facilities and its suicide prevention efforts. It implies that media can have a preventive effect on people with suicidal thinking when they give positive coping skills or emphasize other answers to unfavorable life circumstances (Holding 1975).

More recently, the causes of the protective effect of media have been the subject of intense debate within the scientific community. While some researchers contend that raising awareness of mental illness treatment in public can help prevent suicide (Carmichael and Whitley 2019), others contend that negative media coverage of suicides, such as the suicide victim's "non-attractiveness" or the circumstances of the suicidal act, reduces the likelihood of the imitative effect (Phillips 1978). When suicide is depicted poorly in the media, such as a sad, awful, painful, or disfiguring act, or when suicidal people are portrayed as deviants, the risk of imitation is shown to be reduced. For instance, the lack of the copycat effect is connected to unfavorable media coverage of the 1978 Jonestown mass suicide (Stack 1983) and Kurt Cobain's passing in 1994 (Jobes et al. 1996). In another case, while Romer, Jamieson, and Jamieson observe that newspaper and television news are linked to a rise in suicide fatalities, particularly among younger persons aged 15–25 and persons older than 44 years, the results also show that it has a protective impact among those aged 25–44 (Romer, Jamieson, and Jamieson 2006). In 2010, Niederkrotenthaler et al. discover that reports of people who considered suicide but afterward dealt with their problems constructively are linked to a short-term reduction in suicide rates. This possible preventive effect is coined the "Papageno effect". Inspired by the Papageno effect, Till et al. conduct a randomized controlled trial to explore the beneficial impact of educative newspapers featuring suicide experts in 2018. The results show that readers with or without personal experience of suicide appear to have similar suicide-protective effects. To further test the Papageno effect, in 2022, Niederkrotenthaler et al. conduct a meta-analysis and provide new evidence supporting the beneficial effect of media stories about hope and recovery from suicidal crises on individuals with suicidal ideation.

So far, however, the Papageno effect on social media is understudied. This work

Table 1: Summary of media effects on suicide

Study	Year	Media	Effects Findings
Werther Effect			
Phillips	1978	Newspapers	Stories of murder-suicide lead to other murder-suicides, some of which are misrepresented as airplane accidents.
Stack	1983	Newspaper	The monthly rates of suicide is strongly correlated with the amount of suicidal stories given to the publicity.
Stack	1990	Newspaper	Non-celebrity suicides are linked to rises in the country's suicide rate.
Biblarz et al.	1991	Movies	After seeing films on violence and suicide, the appropriateness of suicide and their emotional arousal levels increase.
Tousignant et al.	2005	Newspapers	It demonstrates an increase in suicide rates shortly after the suicidal reports, particularly by hanging as in the first instance
Kumar et al.	2015	Social Media	By comparing the posting activity and post content after the suicide of ten celebrities, rising suicide thoughts are observed in the post's content.
Papageno Effect			
Niederkrotenthaler et al.	2010	Newspapers	Suicide rates are adversely correlated with coverage of individual suicidal thoughts that are not followed with suicidal behavior.
Arendt, Till, and Niederkrotenthaler	2016	Newspapers	The awareness materials highlighting effective ways to handle a suicide crisis may help people become more connected to life.
Till et al.	2018	Newspapers	Newspapers that provide advice on coping with suicidality, whether they are written by professionals with or without personal experience with suicidality, have suicide-protective effects.
King et al.	2018	Television	By watching television about exploring the relationship between masculinity, mental health, and suicide, subjects are more willing to seek assistance,to refer a friend in need of help, and less willing to adherence to male standards.
Braun et al.	2021	Newspapers	Teenagers seem to benefit from hearing peers' personal accounts of overcoming suicidal thoughts and seeking help in suicide prevention agencies..

attempts to examine the psychosocial impacts of the Papageno effect on Twitter. It gathers longitudinal social media data and compares multiple psychosocial outcomes of individuals engaging in coping story posts with a matched Control group.

2.3 Mental Health and Psycholinguistics

Although suicide is not an inevitable outcome of any psychiatric disease, research shows a link between mental disorders and suicidal behavior. A recent study categorizes the risk factors for suicide into four groups: personality and individual characteristics, cognitive factors, social variables, and adverse life events (O'Connor and Nock 2014). According to a psychiatric autopsy study (Cavanagh et al. 2003), over 90% of suicide victims suffer mental disorders before they attempt suicide. In the early stages of bipolar illness, the risk is very significant. Patients who experience anhedonia and sleeplessness with major anxiety symptoms, alcohol abuse, or emotional problems due to continually shifting emotions have the highest risks for suicide in the near future (Kleespies and Dettmer 2000). Since trait anxiety has a more vital link with suicide risk than state anxiety, the duration of anxiety symptoms is crucial. Nock and Kazdin discover that in predicting outcomes linked to suicide, cognitive characteristics connected with depression are more significant than the emotional component of sadness (Nock and Kazdin 2002).

The task of understanding suicidal ideation attracts researchers from different perspectives. Research suggests that psychological linguistic metrics may be used to understand why people have suicidal ideation (Berman 2005). Lester find the linguistic correlation between suicide and depression and the association between first-person pronouns and physical terms by analyzing a diary left behind by a woman who committed suicide (Lester 2004). The emotional language patterns in the suicidal notes provide a new avenue of research by questing for severe depression and suicide indicators. In another research, real suicide notes are compared to those taken from controls, and the results show that real suicide notes have distinctive characteristics, fewer sentences and words, a decrease in the frequency of certain words, an increase in the use of non-word characters, personal pronouns, verbs, and present-tense words that are not third-person singular (Kim et al. 2019). Rude, Gortner, and Pennebaker discover that people with depression use less positive emotion terms and slightly more negative emotion words than those who had never had a depressive episode. In 2001, Stirman and Pennebaker publish a paper in which they compare the linguistic expression of poets between a small group of suicidal and non-suicidal individuals, using a computerized text-analysis program named Linguistic Inquiry and Word Count (LIWC) (Pennebaker, Francis, and Booth 2001). The program contains a powerful word counter with an internal dictionary, which matches the target words in posts based on emotional tone, linguistic features, personal concerns, and so on. By applying this method, the authors find that compared to poets who did not commit suicide, those who used self-references through first-person singular pronouns, more words related to death, and fewer social references in their texts as they approached the end of their lives (Stirman and Pennebaker 2001). Following that, other studies use LIWC and similar language analysis techniques to

analyze lexical and linguistic features in the text of suicidal individuals from different cultural backgrounds (Fernández-Cabana et al. 2015; Handelman and Lester 2007; Lester, Haines, and Williams 2010). Online language samples from numerous people known to have committed suicide are used in large-scale investigations. One such study, using the Russian edition of the LIWC lexicon, observe that compared to texts from a control group, blog entries written by verified suicidal people include more negative expressions, less social and perception-related lexicon, fewer positive emotion words, and more negative emotion words. These findings provide important light on the linguistic traits of those who committed suicide or are at risk of doing so, as well as the psychological processes behind suicide (Litvinova et al. 2017).

All together, these studies provide a core understanding of leveraging mental and psycholinguistic cues for understanding the Papageno effect on social media. Based on the public content shared on social media platforms, this work focuses on inferring psychosocial outcomes from the perspectives of affect, behavior, and cognition.

2.4 Mental Health, Suicide, and Social Media

The emergence of social media provides research with a new powerful "lens" to give insights into suicide related behavior. Social media reduces the barriers to reaching vulnerable populations, such as individuals with suicidal ideation, to engage in online activities, such as seeking out social support, exchanging information, and maintaining social relationships (Steinke et al. 2017). Prior research uses social media posts to investigate mental disorders (De Choudhury et al. 2013), risk suicide behavior (Coppersmith et al. 2018), and other mental health concerns (Xiang et al. 2021; Verma et al. 2022). Social media provides present and post notes or messages of mental health status, which reduces the bias caused by retrospective analysis and allows for timely intervention or even the prediction of high-risk suicidal behaviors.

Researchers utilize social media data to examine mental health disorders and other suicide factors on social media. According to a systematic study (Marchant et al. 2017), when using social media, teens are more likely to vent their emotions online than to an adult or peer. The research look at words made by at-risk South Korean teenagers on social media and find that an average of 19% of teenagers indicate suicidality online, with the most prevalent causes of academic performance, self-image, bullying, and health issues (Song et al. 2016). In 2013, De Choudhury et al. show the potential to use social media data to measure and predict major depression among Twitter users. This study find that users with depression show less interaction, greater negative emotions, and higher use of first-person pronouns (De Choudhury et al. 2013).

More recently, De Choudhury et al. demonstrate the importance of linguistic features on social media posts to predict users who move from mental health discourse to suicidal ideation. Based on a small subset of Reddit postings, the authors identify various linguistic markers that describe the shifts to suicidal ideation, including social interaction, hopelessness, anxiety, and impulsiveness (De Choudhury et al. 2016). An empirical study investigates Twitter users who represent suicide attempts in the past and analyzes the language and emotions in tweets posted before and after

their suicide attempts. Interestingly, they observe a rise in the proportion of tweets expressing grief in the weeks before a suicide attempt, which follows a pronounced surge in anger and sadness in the week after a suicide attempt (Coppersmith et al. 2016). O'dea et al. demonstrate that individuals use Twitter to discuss their thoughts about suicidal ideation, and it is possible to determine the level of anxiety among tweets about suicide using both human coders and an artificial machine classifier (O'dea et al. 2015).

Relatedly, Kumar et al. compare the posting activity and content following celebrity suicides to find a rise in posting frequency and increased suicidal ideation (Kumar et al. 2015). Another work reveals the importance of using linguistic features to predict users who move from mental health discourse to suicidal ideation (De Choudhury et al. 2016). Based on Reddit postings, the authors develop a propensity score matching to investigate how individuals may discuss their suicidal ideation while controlling for the previous use of linguistic features of mental health. Following this work, De Choudhury and Kıcıman utilize a similar matching approach to study the effect of social support on the risk of suicidal ideation (De Choudhury and Kıcıman 2017).

This work draws motivation from the above body of work in examining the prevalence of the Papageno effect following being exposed to coping story posts on suicidal ideation on social media. It adopts natural language processing and causal inference analyses to provide a computational framework for measuring this effect and reveals important insights about how people show changes in social media behaviors after engaging in coping story posts.

2.5 Suicidal Ideation Classification

Due to the social stigmatization, cultural discrimination, and potential financial issues faced by people with suicidal ideation, people suffering from suicidal ideation might not seek help from the people surrounding them or consult psychiatrists. With the increasing suicidal rates in the world, Natural language processing and machine learning have recently drawn considerable attention from many researchers to investigate the link between suicidal ideation from individual-generated content.

Individual online texts contain rich emotions and linguistic features. Through exploratory the content of social media posts, it is possible to gain insight into the language use and linguistic indicators of suicidal ideation. Ji et al. use syntactic, linguistic, word embedding, word cloud, and topic modeling on suicide related text and reveal that people with suicidal ideation express strong negative feelings, anxiety, and hopelessness and discuss personal and social issues (Ji et al. 2018). Another work summarizes nine topics in the text of people with suicidal ideation and manually extract extracts a set of keywords related to these topics (Abboute et al. 2014). Shing et al. develop a dataset using social media posts on a discussion forum called r/*SuicideWatch* on Reddit. They extract a variety of features, such as a bag of words, empath, readability, syntactic features, topic model posteriors, word embeddings, LIWC lexicon, emotion features, and mental disease lexicon, for automatic risk-level classification.

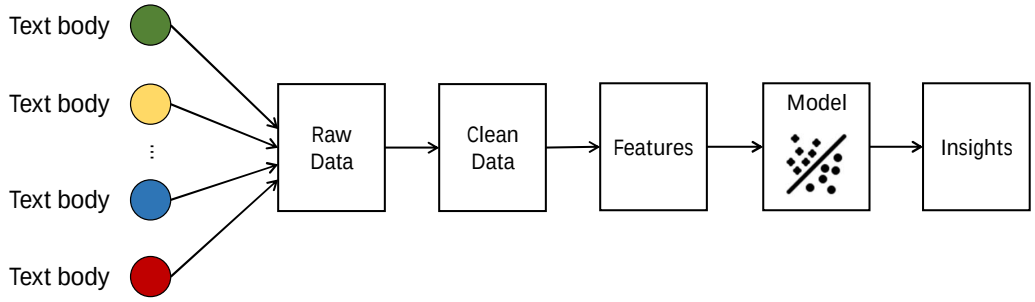


Figure 3: The illustration of utilizing features in text to build machine learning classifier

Classifying suicidal ideation from the text can be cast as a text classification, where NLP techniques are used to extract features and machine learning algorithms are applied for the tasks of classifications (Figure 3). Burnap et al. extract lexical, structural, emotive, and psychological features from Twitter posts and build several machine learning classifiers, including Support Vector Machine (SVM), Decision Tree, and Naïve Bayes, to identify suicidal ideation in Twitter posts (Burnap et al. 2017). Using both human annotators and machine learning classifiers based on frequency-inverse document frequency (TF-IDF) characteristics, O'dea et al. demonstrate the possibility of categorizing the concern level among Twitter posts related to suicide (O'dea et al. 2015). Using machine learning algorithms, Braithwaite et al. show the efficiency of distinguishing individuals at suicidal risk and individuals not at suicidal risk. Sawhney et al. improve the ability of the Random Forest (RF) classifier to detect suicidal ideation in Twitter posts.

Current machine learning methods integrate neural network models for detection of suicidal ideation. With well-known word embedding methods like word2vec (Mikolov et al. 2013a) and GloVe (Pennington, Socher, and Manning 2014), natural language text is often embedded into distributed vector space. Tadesse et al. propose a deep learning model and multi-class machine-based approach for the detection early signs of suicidal ideation on social media posts. By comparing the Convolutional Neural Network (CNN), Long short-term memory (LSTM), and LSTM-CNN combined model, they demonstrate the potential use of applying the combined neural network model with word embedding techniques into suicidal post detection. Another research implements a deep neural net (DNN) based on the psycholinguistic feature and word occurrence. They compare the difference between suicidal and depression periods in the daily content of their text messages (Nobles et al. 2018). Metzler et al. train two machine learning classifiers and two deep learning models (Bidirectional Encoder Representations from Transformers (BERT) and XLNet) to classify Twitter posts into six suicidal related categories. They show deep learning classifier performances are close to human performance in most cases.

This study is inspired by the above body of work in classifying suicidal ideation. This work utilizes a machine learning classifier from a previous study (Metzler et

al. 2021) to automatically annotate the dataset. A sample of the machine-labeled dataset is manually annotated to compare the machine-labeled and human-labeled data to verify the accuracy of the classifier.

3 Datasets

Due to the absence of publicly available datasets of coping story posts in social media, this work utilizes Twitter timeline data of individuals who reply to coping story Twitter posts. The steps of data collection include: 1) collecting Twitter posts, which might describe coping stories. 2) applying a coping story classifier from (Metzler et al. 2021) and manually verifying the results; 3) collecting timeline data of individuals commenting on the so found coping story posts; 4) building a Control dataset from randomly sampled, comparable, individuals.

3.1 Twitter Terminology

This study focuses on Twitter, a popular social media where people communicate in short posts called *tweets*. A Tweet typically consists of photos, videos, links, and text. Twitter users data contain their metadata, including unique id, unique username and account creating time, and timeline data – a collection of Tweets posted by the users. Users on Twitter have four different interactions to connect with other users: follow, like, retweet and reply.

- **Follow**: Follows is a directed relationship where a user subscribes to view the content (tweets) posted by another user.
- **Retweet**: Typically, a retweet signifies an endorsement of the topic of the tweet. In certain instances, though, it may reflect a different explanation, such as someone making fun of or criticizing the post.
- **Reply**: One of the simplest ways to participate in a discussion on Twitter is to respond to another Tweet, which is what the term "reply" refers to.
- **Like**: Likes are a common way to interact with other Twitter posts. They indicate approval or upvote by using a little heart.

3.2 Data Sources

This study focuses on the prominent social networking website Twitter. Twitter enables registered users to publish and read communications known as tweets. This work also considers retweeting, the act of sharing a tweet that was previously made by another user, to be a user activity. This study uses Twitter academic API [2] and its fullarchive search is utilized to get the data for scholarly research purposes. This API offers programmatic access to public Tweets from the entire archive, dating back to the first Tweet in March 2006.

Three different search methods are undertaken on Twitter for this research. The broad Twitter posts search is utilized to build a list of people whose Twitter posts contain specific keywords or phrases. The returned results have the user's unique id, twitter texts, and posting timestamp. Next, this project uses Twitter user search

[2] https://developer.twitter.com/en/products/twitter-api/academic-research

Table 2: Paraphrased example Twitter posts labeled with coping story or non-coping story.

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 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."

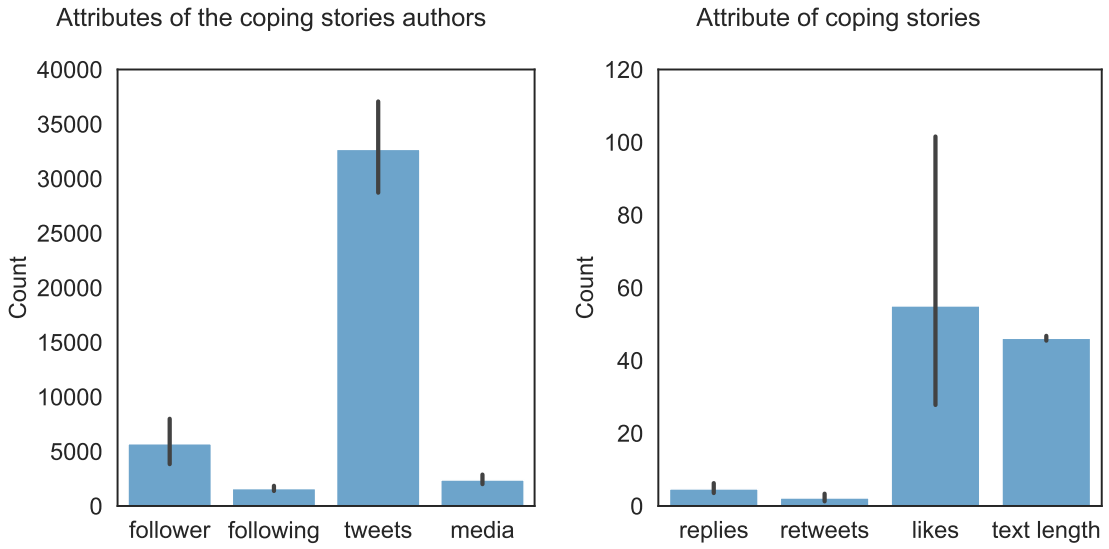
to collect timeline data, which provides access to tweets made by a specific Twitter account. User lookup is utilized to obtain user information like account creation date, the number of posts, location, and name.

3.3 Compiling Coping Stories Dataset

The first step is to retrieve Twitter posts that may contain coping stories. This work uses the Twitter Application Programming Interface (API) to collect Twitter posts posted between 1 January 2018 and 1 March 2022. The collected Twitter posts contain at least one term related to suicide attempts, such as "suicide thoughts", "kill myself" and "end my life", and contain at least one of the terms indicating successful coping, such as "happier", "better" and "recover". After collecting 13,022 Twitter posts, the project applies the coping story classifier to annotate each Twitter post as a coping story or a non-coping story. This work finds 3,077 Twitter posts are annotated as coping story posts (Figure 4). Among them, 709 Twitter posts labeled coping story posts have at least one reply below them.

3.3.1 Annotating Coping Story posts

As shown in Figure 5, a multi-label classifier provided by Metzler et al. is used to annotate coping story posts in the dataset. It categorizes Twitter posts into the following six categories: personal story of either suicidal ideation or suicide attempt, coping story, call for action intending to spread either problem awareness or prevention-related information, report of suicide cases, and irrelevant Twitter post. Metzler et al. use the term "coping story" to refer to *"personal stories about*



(a) (b)

Figure 4: (a) Attributions of coping stories author, (b) Attributions of coping stories.

an individual's experience with suicidal thoughts or a suicide attempt, with a sense of hope, recovery, coping, or mentioning an alternative to suicide" (Metzler et al. 2021). As the aim of this step is to find individuals who comment on coping story posts, this work subsequently focus on Twitter posts labeled as a coping story post. In order to verify the ability of the classifier to find Twitter posts that contain actual coping stories, the author manually checks Twitter posts that are labeled as coping story posts by the classifier.

3.3.2 Annotation Task

400 Twitter posts are randomly sampled out of 709 Twitter posts labeled coping story that have at least one reply below them. Using the codebook from (Metzler et al. 2021), The author of this thesis independently annotates 400 Twitter posts. If there are any posts that they are unsure about, the author discusses the posts with other a collaborator and together they agree on how to code them. After finishing the annotation, the collaborator randomly selects 50 posts out of 400 posts to verify the annotation result. Cohen's kappa is used to validate the annotation process. This results in Cohen's k of 0.81 with an agreement of 92.8%, which suggests a substantial agreement (Landis and Koch 1977). Out of the 400 posts labeled coping story by the classifier, the work find 347 posts are correct predictions, indicating 86.7% accuracy of the classifier to identify a copy story post.

3.4 Compiling Treatment Users Dataset

For the Twitter posts annotated as coping story posts, this study assumes that the coping story might have impacted the individuals who reply to the Twitter posts. For

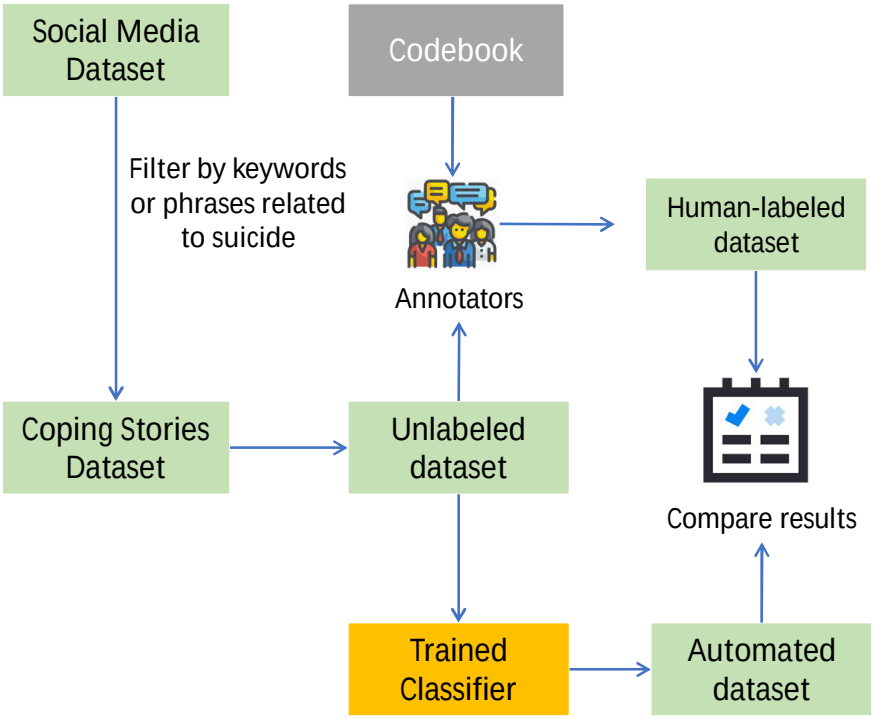


Figure 5: Annotation workflow.

the coping story dataset, 2470 unique individuals who reply to at least one coping story are identified. The project collects Twitter metadata, including the number of posts, likes, followers, followees, and the account creation time. Two individuals who reply to multiple coping story posts are removed to mitigate the confounding effects of engaging in multiple coping story posts. The remaining 2468 individuals' timeline data are two weeks before their reply on the coping story and two weeks after being collected (Figure 7a). In the end, the target dataset contains 2468 individuals with a total of 787K Twitter posts.

3.5 Compiling Control Users Dataset

As the goal of this project is to seek to isolate the effect of the coping story on individuals, this project builds a Control dataset of individuals who do not reply to coping story posts during the investigated period. To do so, a Control dataset is built with individuals who have similar attributes to the Treatment individuals prior to their engagement with a coping story. To find such Control individuals, this work uses keywords, such as "life", "job", "music", and "movie" to search for individuals on Twitter. For each keyword, the timelines of Twitter posts from 4000 individuals are collected. In the end, For each individual in the Control dataset, a placebo date from the non-parametric distribution of Treatment date of the Treatment dataset is assigned to any day the Control individual replies to other Twitter posts to reduce

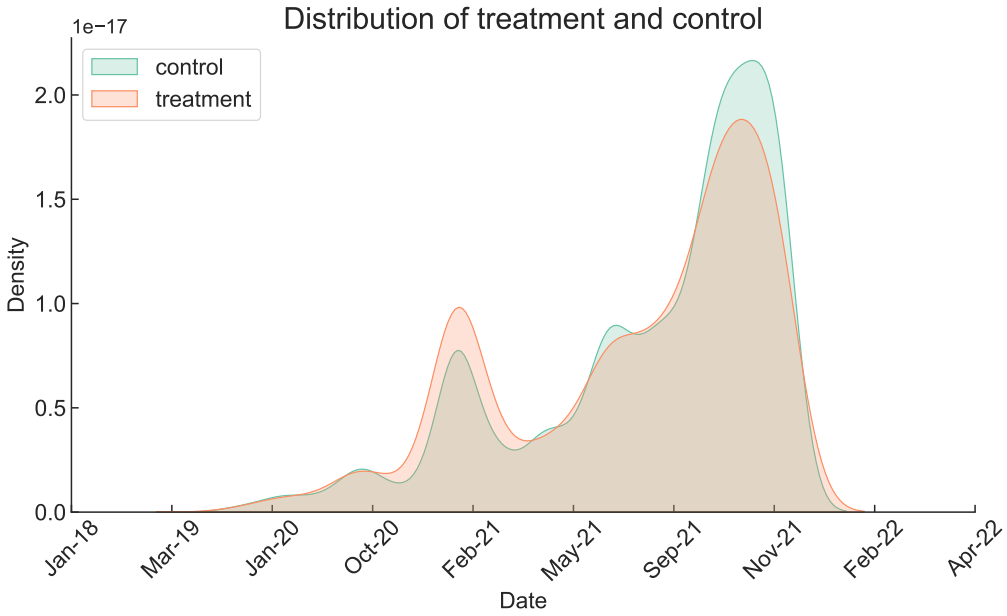
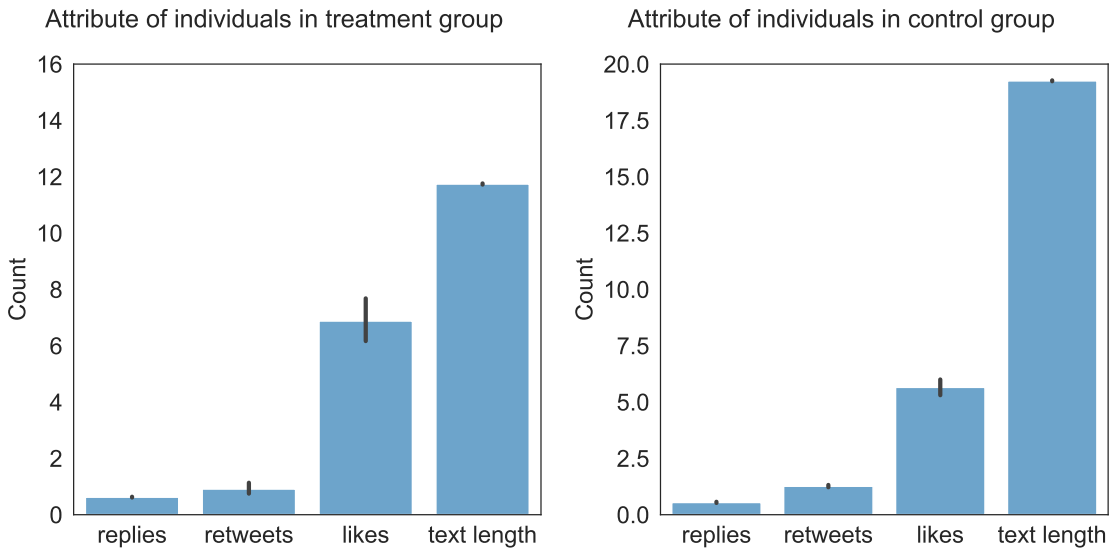


Figure 6: Treatment and Control (placebo) dates distribution.



(a) (b)

Figure 7: (a) Attributions of Twitter posts in Treatment groups before and after 2 weeks replying to coping stories, (b) Attributions of Twitter posts in Control groups before and after 2 weeks replying to coping stories.

any temporal confounds. Kolmogorov-Smirnov (KS) test is utilized to measure the similarity in the two distributions (Figure 6). The KS test yields a low statistic of 0.05, suggesting that the probability distributions of treatment and placebo dates are similar. In the end, By limiting the timeline data two weeks before the placebo date and two weeks after, there are 8465 individuals in the Control dataset.

3.6 Preprocessing of the Data

Data preprocessing is essential for enhancing social network analysis performance. Social network data often includes text, emoticons, URLs, and other diverse types of data. The datasets in this project are cleaned using the procedures listed below.

- **Punctuation Replacement** This project removes all the punctuation from the text in the Twitter posts. Punctuation is considered non-vital information in linguistic analysis.
- **Stop word elimination** Special characters like "%," "@," and "#"; as well as useless acronyms like "AFK" and "IDK" are included in the stop words. As these words do not significantly contribute to the meaning of the text, they are removed before the text processing.
- **Emojis emoticons** Emoji is a picture word encoded by characters and emoticons. The emojis in the text are removed as they can not be analyzed by linguistic analysis. Further studies might use emojis to deeply investigate the sentiment change after engaging in coping stories.

4 Method

4.1 Study Design and Rationale

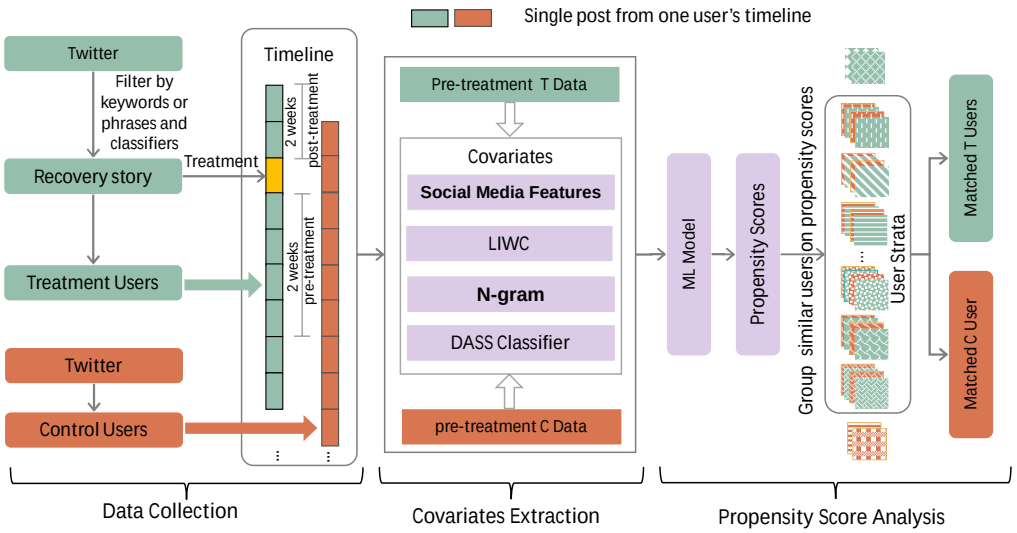


Figure 8: Schematic illustration of stratified propensity score framework between Treatment individuals and Control individuals.

This project adopts a stratified propensity score framework (Imbens and Rubin 2015) to isolate the Papageno effect. The schematic illustration of this methodology is shown in Figure 8. This method first matches individuals of the Control group with individuals of the Treatment group by controlling pre-Treatment behavioral attributes. To do this, a machine learning classifier is trained to predict the probability of an individual being allocated to either the Treatment or Control group (i.e., *propensity*) based on covariates and execute matching across groups using estimated propensity scores. Within matched groups of Control and Treatment groups, the following psychosocial outcomes are analyzed between matched Control groups and Treatment groups. In sum, this approach ensures that members of the Treatment group and Control group who are being compared have similar behavior prior to replying to coping story posts. This provides us with the tools to compare the differences in psychosocial outcomes between the matched Treatment individuals and the members of the Control group.

4.2 Measuring Psychosocial Outcomes

To achieve the study aim of understanding the Papageno effect on social media, the improvements in psychosocial health outcomes are quantified on Twitter. Psychosocial health involves psychological and social wellbeing, and places the center of health inside the person by incorporating social wellbeing in the form of social adjustment and environmental reaction (Larson 1996). This work is situated in terms of the

psychological health and wellbeing effects of social media interventions (Merolli, Gray, and Martin-Sanchez 2013). A conservative definition of psychosocial health is employed based on observed behavior on social media platforms due to the fact that psychosocial health is a complicated concept for which there is no straightforward method of measurement. As individuals' posting behavior is the only accessible data, three psychosocial consequences are quantified based on findings from a study of mental health and social media (Saha, Weber, and De Choudhury 2018a; Saha and Sharma 2020), grounded on linguistic, behavioral, and psychology literature. These psychosocial outcomes are broadly grouped into three categories: affective, behavioral, and cognitive outcomes (Breckler 1984).

Affective Outcomes Affect is defined as any experience of feeling or emotion (VandenBos 2007). As online individuals use emotive, relativistic language in their self-motivated texts, language is an effective way to infer affective psychosocial wellbeing. To estimate affective outcomes, this work uses the following metrics:

Affective Words. The well-validated psycholinguistic lexicon, Linguistic Inquiry and Word Count (LIWC) (Pennebaker and Chung 2007) is employed to obtain normalized occurrences of words in affective categories (positive emotion, negative emotion, anger, sadness) per individual. The selection of these measures is inspired by studies like (Ernala et al. 2017; De Choudhury et al. 2013; Saha, Weber, and De Choudhury 2018b) where therapeutic symptoms are associated with self-initiated and expressive writing (Cohn, Mehl, and Pennebaker 2004; Chung and Pennebaker 2007).

Symptomatic Mental Health Expressions. Prior research notes the comorbidity of multiple mental health conditions (Rosenblat, Kakar, and McIntyre 2016), and the linguistic indicators of different mental health symptomatic expressions of depression, anxiety, stress, and suicidal ideation are operationalized (Saha et al. 2019). Saha et al. create a variety of binary machine learning classifiers based on transfer learning techniques to recognize mental health symptomatic expressions in social media language (Saha et al. 2019). These classifiers are n-gram–based (n=1,2,3) binary support vector machine (SVM) models, and are trained using appropriate Reddit communities (*r/depression* for depression, *r/anxiety* for anxiety, *r/stress* for stress, and *r/SuicideWatch* for suicidal ideation). People post in these communities about mental health symptoms to receive feedback and to support others. The posts in these subreddits are used as training data to identify language used in connection with mental health. The training data for texts not related to mental health originates from non-mental-health-related content on Reddit. These classifiers perform at high average accuracy of 0.90 on test data (Saha et al. 2019), and have also been used in other research (Saha and Sharma 2020; Saha et al. 2020). The study uses these classifiers to measure the aggregated proportion of expressing mental health concerns per individual. A Lower quantity of posts on mental health symptomatic expressions indicates a better psychosocial wellbeing.

Behavioral Outcomes. Literature in psychology defines behavioral psychological health into three factors: An individual's overt actions, behavioral intentions, and a verbal statement regarding behavior (Breckler 1984). Previous studies quantify behavioral psychological wellbeing by measuring the shifts in social functioning and

interests (Guntuku et al. 2019; Saha, Weber, and De Choudhury 2018a). This work operationalizes the following measures to obtain behavioral outcomes.

Activity. The study is interested in investigating if engaging in coping story posts promotes individuals to be more active on Twitter. Higher activity likely indicates increased extroversion, and is associated with therapeutic benefits (Ernala et al. 2017; Saha, Weber, and De Choudhury 2018a). To quantify activity on Twitter, the average number of Twitter posts per day are calculated for every individual. The higher activity score means better psychosocial outcomes.

Interactivity. Interactivity is another indicator of an individual showing therapeutic effects (Saha, Weber, and De Choudhury 2018a; Saha and Sharma 2020). The participation in discussions on Twitter is measured as interactivity, indicating social engagement. The metric used is the proportion of replies (to other individuals' posts) per original Twitter post. The higher interactivity score indicates an improvement of social engagement.

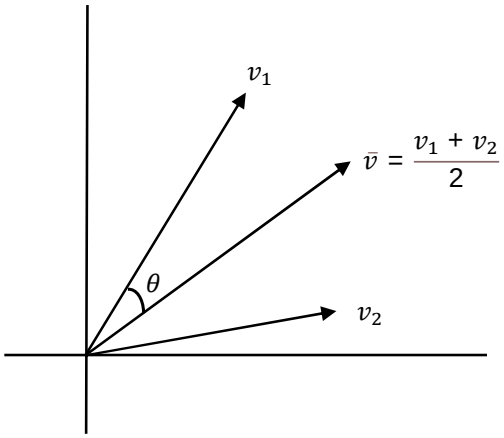


Figure 9: Vectors of word embeddings models

Topic Diversity. To measure the diversity of topics discussed by an individual, a language model on the posted texts is deployed. The language semantics are captured by adopting a word embedding model (Mikolov et al. 2013b). It represents words in vectors in latent semantic dimensions. In particular, 300-dimensional word embeddings pre-trained on Google News are used in this study. As shown in Figure 1, for each post in the Treatment and Control datasets, the average similarity score (cosine distance) from Twitter posts to the centroid of the corresponding corpus is calculated by

$$\text{similarity} \ = \cos(\theta) = \frac{\hat{\mathbf{v}} \cdot \mathbf{v_i}}{\|\mathbf{v}\|\mathbf{v_i}\|} \tag{10}$$

After that, the average similarity for every individual is estimated in the the Treatment and Control datasets. The higher the average the distance from the centroid, the greater is the diversity of expressed topics in posts (Wang et al. 2021).

Cognitive Outcomes. The cognitive component of psychosocial health consists of beliefs, knowledge structures, perceptual responses, and thoughts (Breckler 1984). The following measures are adopted to quantify an individual's cognitive behaviors.

Readability. Readability measures the complexity of a given text. The Coleman-Liau Index (CLI)is adopted to assess the readability per individual. CLI is calculated as,

$$CLI = 0.0588 * L - 0.296 * S - 15.8, \tag{11}$$

where L represents the average number of letters in a word (per 100), and S represents the average number of sentences in a word (per 100) (Kher, Johnson, and Griffith 2017). Previous research shows a link between measures of language complexity and long-term improvements in psychosocial wellbeing (Ernala et al. 2017). The higher readability score, the higher psychosocial wellbeing.

Complexity and Repeatability. Complexity and repeatability are syntactic measurements that reflect an individual's cognitive state in terms of planning, execution, and memory (Ernala et al. 2017). This study measure repeatability as the normalized count of non-unique words and complexity as the average number of words per sentence. Language complexity and psychosocial wellbeing are positively correlated, while repeatability is negatively correlated with psychosocial wellbeing.

Psycholinguistic Keywords. Typically, online user postings use emotive, relativistic, and harassing language. Lexicons are extensively used to extract these characteristics. This study uses the LIWC lexicon to analyze the proportion of keywords related to cognition, perception, social context, and linguistic style categories. This study considers the following five aggregated categories:

- **Cognition & Perception**: *cause, certain, cognitive, inhibition, discrepancies, tentativeness, perception, see, hear, feel, insight.*
- **Social Context**: *biological processes, achievement, body, family, friends, health, home, humans, money, religion, social, work.*
- **Lexical Density & Awareness**: *adverbs, article, verbs, auxiliary verbs, conjunctions, inclusive, exclusive, preposition, negation, quantifier, relative.*
- **Interpersonal Focus**: *1st personal pronouns, 2nd personal pronouns, Impersonal pronouns.*
- **Temporal References**: *future, past, present.*

Prior research highlight the association between these lexicons and cognitioni (Pennebaker, Mehl, and Niederhoffer 2003). Increased use of these lexicons is related to better psychological conditions.

4.3 Matching For Causal Inference

4.3.1 Matching Covariates

This aim of this study is to measure the psychosocial outcomes of engaging in a coping story post. Satisfied propensity score matching is an efficient strategy in Case-

Controlled studies to measure causal effects and to minimize the possible occurrence of selection biases. In this case, several covariates are extracted from the Treatment and Control datasets to control for similar pre-Treatment behavior on social media. In the context of this study, several covariates are built from the target and Control datasets to control for similar pre-Treatment behavior on Twitter.

- **Social Media Feature:** The first set of covariates consists of individuals' social media features, including the number of likes, the number of followers, the number of followees, and posting frequency)
- **Unigrams:** The distribution of words used in their Twitter timelines is included in the second set of covariates. This work extracts the top 100 unigrams as the second covariates set.
- **Linguistic Features:** By examining the word distribution in the LIWC lexicon, the third set measures the average usages of psycholinguistic lexicons across all Twitter posts during the pre-Treatment period.
- **Depression, Anxiety, Stress, Psychosis, Suicidal Ideation:** the average number of Twitter posts containing symptomatic mental health expressions, including depression, anxiety, stress, and suicidal ideation.

4.3.2 Logistic Regression For Propensity Score Estimation

The majority of propensity score applications adopt logistic regression to estimate the score (Austin 2011). Logistic Regression (LR) is a supervised machine learning algorithm, that works by determining the output variable probabilities. In the context of propensity score matching, Treatment assignment is used as the dependent variable in LR, while all selected covariates are used as independent variables. Estimating the log probability of an event is the main job in the logistic regression analysis. Let T_i be a binary Treatment/Control indicator for the ith subject. Thus, for each subject, the propensity scores are estimated; that is, the conditional probability of being addicted given the observed covariates X_i. Logistic regression mathematically calculates a multiple linear regression function that is defined as:

$$\log \frac{P\left(T_i = 1 \mid X_i\right)}{1 - P\left(T_i = 1 \mid X_i\right)} = \alpha + \beta^T X_i \tag{12}$$

4.3.3 Propensity Score Analysis

To control the similarities between the Treatment dataset and the Control dataset, satisfied propensity score matching is used to pair Treatment users and Control users, which control pre-Treatment covariates are similar to each other. After a logistic regression classifier is implemented to estimate the likelihood of a user belonging to the Treatment group or Control group based on their covariates, the propensity score distribution is divided into 50 strata with equal lengths. Individuals with similar propensity scores are assigned into the same stratum (Kiciman, Counts, and Gasser

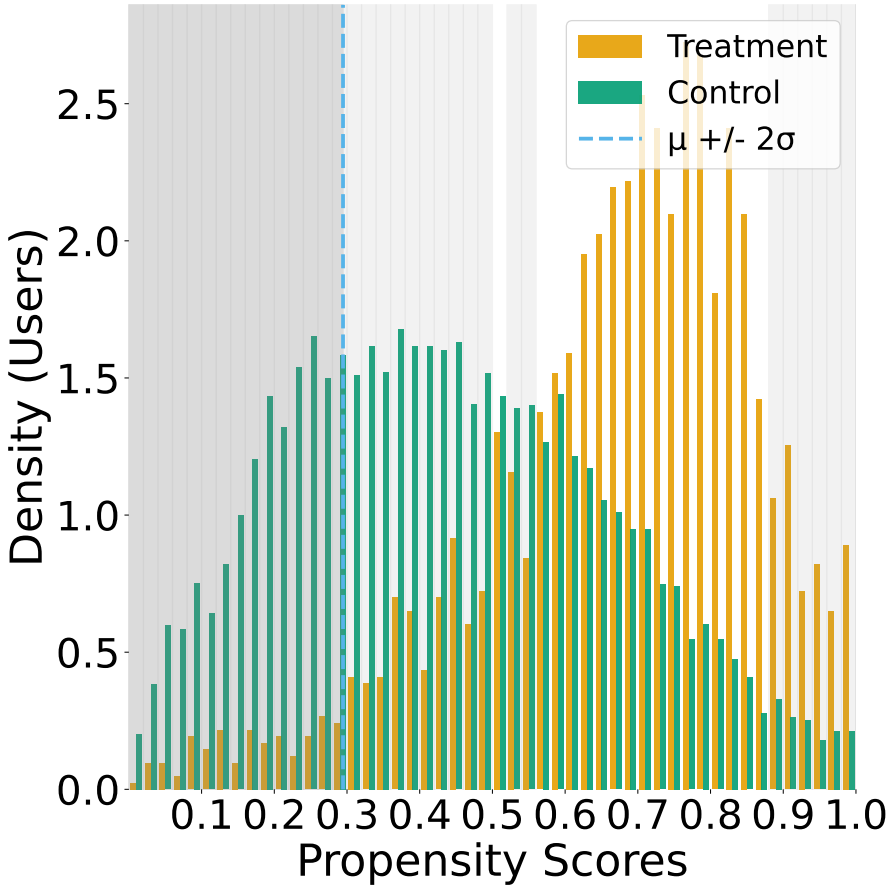


Figure 10: Propensity score distribution (shaded region are the dropped strata)

2018). This aids in evaluating possible psychosocial outcomes within each stratum, where the Control group individuals are matched to the Treatment individuals based on the pre-Treatment behavioral traits. This study removes the individuals with propensity scores falling within two standard deviations from the mean (Figure 10). The strata are dropped if it fails to satisfy the minimum sample size within each stratum based on previous causal inference research (De Choudhury, Sharma, and Kiciman 2016). By limiting that there are at least 50 individuals per group in each stratum, this approach yields in 14 strata with 1245 Treatment and 1087 Control individuals.

4.3.4 Quality Assessment of Covariate Matching

To assess whether individuals in the Treatment group and Control group are statistically comparable, the balance of the covariates are estimated. This comparison are conducted by calculating the standardized mean differences (SMD) between the two groups in all the 14 valid strata (Saha et al. 2019; Kiciman, Counts, and Gasser

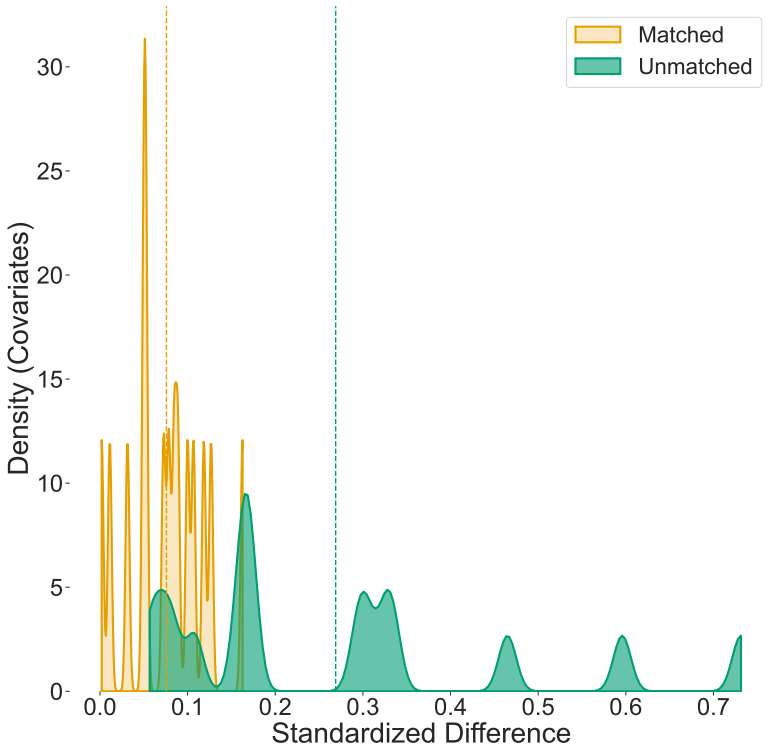


Figure 11: Quality of matching

2018). For each outcomes, the SMD is calculated within matched groups by

$$SD_{\text{pooled}} = \sqrt{\frac{(SD_1^2 + SD_2^2)}{2}}, \tag{13}$$

where:

- SD_1 = standard deviation for Treatment group.
- SD_2 = standard deviation for Control group.

The two groups can be assumed to be balanced if the SMD of all covariates is lower than 0.2 (Kiciman, Counts, and Gasser 2018; Stuart 2010; Saha et al. 2021a).For the unmatched dataset, the maximum SMD is 0.85, and the mean SMD is 0.22, whereas in the matched dataset, the maximum SMD is 0.19 and the mean SMD is 0.06 (Figure 11). Therefore, this satisfies the threshold of SMD<0.2, suggesting that the matching yields balanced Treatment and Control datasets.

4.4 Estimating the Average Treatment Effects

To estimate the effect of engaging with coping story posts, this work computes the relative treatment effect (RTE) for each outcome. For this, this project calculates

the ratio of the possibility of an outcome measure in the Treatment group to that in the Control group per stratum. Using the number of Treatment individuals in each stratum as a weight, the weighted average RTE per outcome is calculated. The outcome is interpreted as an increase (greater than 1) or decrease (less than 1) of observable psychosocial outcomes after engaging with coping story posts compared to the Control group with similar pre-Treatment attributes.

$$RTE_k = \frac{\sum_i^{N_k} S_{L_i}^{\text{after}}}{N_k} / \frac{\sum_j^{M_k} S_{C_j}^{\text{after}}}{M_k}, \tag{14}$$

where:

- N_k is the number of matched Treatment users in stratum k.
- M_k is the number of matched Control users in stratum k.
- $S_{L_i}^{\text{after}}$ the psychosocial outcomes for matched Treatment user L_i.
- $S_{C_j}^{\text{after}}$ the psychosocial outcomes for matched Control user C_j.

5 Result

In this section, the shifts for each psychosocial outcome are presented across the matched Treatment and Control individuals in the corresponding datasets. This project calculates the effect size (Cohen's d), and measures the statistical significance in differences using an independent sample t-test. Figure 12 shows the distribution of RTE per stratum across different psychosocial outcomes. Table 3 summarizes these differences.

Affective Outcomes. *Affect.* In Table 3, The results observe that individuals in the Treatment group use more affective words than the matched Control individuals after engaging with coping story posts. The average of affective words used among Treatment individuals is 9% higher than among Control individuals. The effect size (Cohen's d=0.20) suggests small differences between the two distributions and the t-test indicates statistical significance (t=2.17, p<0.05). This observation affirms that individuals use more affective words after engaging with coping story posts.

Table 3: Summary of psychosocial differences across all the outcomes between Treatment and Control individuals. The results report mean psychosocial outcomes across all matched individuals, effect size (Cohen's d), independent sample t-statistic ($*p < 0.05, **p < 0.01, ***p < 0.001$).

Categories	Tr.	Ct.	RTE	d	t-test
Affective Outcomes					
LIWC: Affect	0.11	0.10	1.06	0.20	2.17*
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

Symptomatic Mental Health Expressions. The results find that engaging with coping story posts is associated with decreases in using symptomatic stress and depression

expressions. This is revealed by lower average percentages of symptomatic stress and depression Twitter posts from individuals in the Treatment group reflecting stress (t=-3.96, p<0.001) and depression (t=-2.84, p<0.01). In contrast, the results show no significant differences in the measures of anxiety and suicidal ideation after individuals engage with coping story posts between the two corresponding groups. This illustrates that engaging with coping story posts does not increase the use of symptomatic anxiety and suicidal ideation expressions. These observations are consistent with prior research which indicates that media featuring individuals coping with depression and suicidal ideation reduces depression and shows no effect on suicidal ideation (Niederkrotenthaler and Till 2020).

Figure 12: RTE per propensity stratum per outcome. The pink color indicates RTE greater than 1, while the green color suggests RTE less than 1.

Behavioral Outcomes. For the second set of outcomes, the results find no significant difference in activity after engaging with coping story posts. The results shows that the average interactivity of the Treatment users is higher than the Control. Both effect size (Cohen's d=0.34) and independent t-test indicate statistical significance (t=4.17, p<0.01). This might suggest that engaging with coping story posts likely

promotes an individual's participation in online discussions. For topical diversity, this project measure the diversity of expressed topics in posts after engaging with coping story posts; the effect size informs small differences between Treatment and Control distributions of topical diversity, and the t-test affirms statistical significance (t=2.51, p<0.05). This indicates individuals in Treatment group posts discuss broader topics after engaging in coping story posts, suggesting higher psychological wellbeing.

Cognitive outcomes. To examine if engaging with coping story posts leads to shifts in cognition, the project measure the differences in readability, complexity, repeatability, and psycholinguistic features. Among these, the results find no significant differences in complexity, cognition & perception, lexical density & awareness, and temporal reference. However, the results observe a significant difference in interpersonal focus (t=2.77, p<0.05). The changes in the usage of pronouns might suggest a shift in how individuals see themselves in relation to others. Although independent sample t indicates a significant difference in readability (t=-2.17, p<0.05), the result of cohen's d shows no difference between the two distributions. One unanticipated finding is the Treatment individuals show higher repeatability compared to Controlindividual (t=5.09, p<0.001), which suggest lower psychosocial health.

6 Discussion

This study provides a causal inference approach to access the Papageno effect on individuals who engage with coping stories on social media. This section discusses the implications from a theoretical prospective, as well as practical and design implications. Next, the limitations and future works are presented in details.

6.1 Implications

6.1.1 Theoretical Implications

While this study does not directly measure shifts in suicidal ideation itself, the findings still provide evidence for the Papageno effect on social media. The results suggest that engaging with posts describing personal stories featuring coping with suicidal ideation can positively impact psychosocial wellbeing. This work provides a methodology to help measure the psychosocial outcomes of the Papageno effect on social media. By focusing on the psychosocial shifts of a large sample of individuals who engage with coping story posts, this study suggests a role for utilizing social media to access the prospective psychosocial outcomes of the Papageno effect.

This work provides a data-driven methodology to verify the Papageno effect on social media and reveal how the Papageno effect can impact the psychosocial outcomes of individuals. The work bears implications for future studies about mental health and suicide on social media by providing a novel causal inference framework for investigating the psychosocial shifts on social media. Despite the emerging research on the Papageno effect, much less is known about the Papageno effect on social media. Previous research also notes the potential of using large-scale social media data to characterize suicidal behaviors and help-seeking (Metzler et al. 2021). By focusing on the psychosocial shifts of a large sample of individuals who engage with coping story posts, this work suggests a role for utilizing social media to quantify the prospective psychosocial outcomes of the Papageno effect.

6.1.2 Practical and Design Implications

Currently, the majority of suicide related online communities only act as safe networking spaces but are missing guidelines to discuss suicidal ideation or previous suicidal attempts. Platforms such as Reddit, Twitter, or TalkLife rely on individual reports or moderators observations on an online discussion without decision support. This work bears practical implications for developing suicide discussion guidelines. Online communities may develop strategies for the narratives about sharing suicidal behaviors and ideation, allowing vulnerable members in the community to be more protected. This work bears design implications for social media platforms in terms of how these platforms can encourage positive and thriving behavior, especially for those struggling with mental health concerns. Platforms such as Reddit, TalkLife, and 7Cups follow community-driven moderation strategies, and these platforms can include in their community norms what kinds of postings and support can help people draw therapeutic benefits. Social media platforms can show more posts with

the Papageno effect when individuals search for suicide related information. These may be beneficial for individuals engaging with coping story posts and helpful for individuals with suicidal ideation to seek support and prevent tragic outcomes.

The results show that engaging with Twitter posts about coping with suicidal ideation improves expressive writing. Previous research finds that self-motivated disclosure and expressive writing are positively associated with mental wellbeing (Ernala et al. 2017), because online platforms encourage people to share stigmatized experiences, such as mental health and suicidal experience. Accordingly, this study envisions the promotion of a coping stories discussion forum which may be built and integrated directly or indirectly with social media platforms, where people can share and discuss their experiences about how to cope with their suicidal ideation without mentioning special suicidal ideation. It can protect people suffering from suicidal ideation and enable them to express themselves. Moreover, these posts can serve as a document about their suicidal ideation, and reflections on their thoughts and feelings. Previous research indicates that archival writing can improve individuals' ability to form a narrative of complex events, experiences, and mental health challenges and meet self-care and coping goals (Pennebaker and Seagal 1999).

6.2 Limitations and Future Work

This study has limitations, some of which point to intriguing future research directions. This study does not account for passive engagement behaviors. For example, it is unknown if an individual reads a coping story and is affected by it if this individual does not reply to it. Another limitation is the time of the measurements. This study only measure the averaged psychosocial outcomes within two weeks after an individual engaged with a coping story post. However, the psychosocial outcomes may vary over time and show fluctuating results (Bin Morshed et al. 2019).

This study suffers from *selection biases*. It gathers data from those who publicly reply to coping story Twitter posts, which is likely to be influenced by self-selection bias. This project can only collect data from those who are active on social media. This is especially important considering the stigma associated with people having suicidal ideation.

This study notes that the results cannot infer "*true causality*", as the outcomes might have be been influenced by other online and offline factors, e.g., the individual's suicidal behavior history, the length of a coping story post, and the length of the replies. Despite corroboration by a psychiatrist, this study cannot be certain based on Twitter posts that the individual is personally seriously considering suicide. Therefore the symptomatic variables and outcomes need additional clinical validation. Future work could combine social media data together with clinically validated data which could lead to more generalizable results on the Papageno effect and the role of social media.

7 conclusion

This work develops a novel causal inference framework to verify and study the Papageno effect on social media. Using a Twitter dataset with ∼2M posts by ∼10K individuals, this work observes statistically significant psychosocial (affective, behavioral, cognitive) shifts in individuals after engaging with coping story posts. In assessing these psychosocial effects, the causal framework controls behavioral and linguistic covariates across the Treatment and Control groups. This work verifies that engaging with coping story posts positively impacts individuals' stress and depression, and improves expressive writing, the diversity of expressed topics in posts, and interactivity. The results indicate that engaging with coping story posts on social media is associated with positive benefits for one's psychosocial wellbeing.

References

Abboute, A.; Boudjeriou, Y.; Entringer, G.; Azé, J.; Bringay, S.; and Poncelet, P. 2014. Mining twitter for suicide prevention. In *International Conference on Applications of Natural Language to Data Bases/Information Systems*, 250–253. Springer.

Ali, M. S.; Prieto-Alhambra, D.; Lopes, L. C.; Ramos, D.; Bispo, N.; Ichihara, M. Y.; Pescarini, J. M.; Williamson, E.; Fiaccone, R. L.; Barreto, M. L.; et al. 2019. Propensity score methods in health technology assessment: principles, extended applications, and recent advances. *Frontiers in pharmacology* 973.

Arendt, F.; Till, B.; and Niederkrotenthaler, T. 2016. Effects of suicide awareness material on implicit suicide cognition: A laboratory experiment. *Health Communication* 31(6):718–726.

Austin, P. C. 2011. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research* 46(3):399–424.

Berman, A. L. 2005. Forensic psychiatry and forensic psychology| psychological autopsy.

Biblarz, A.; Brown, R. M.; Biblarz, D. N.; Pilgrim, M.; and Baldree, B. F. 1991. Media influence on attitudes toward suicide. *Suicide and Life-Threatening Behavior* 21(4):374–384.

Bin Morshed, M.; Saha, K.; Li, R.; D'Mello, S. K.; De Choudhury, M.; Abowd, G. D.; and Plötz, T. 2019. Prediction of Mood Instability with Passive Sensing. *PACM IMWUT*.

Blumenthal, S., and Bergner, L. 1973. Suicide and newspapers: A replicated study. *American Journal of Psychiatry* 130(4):468–471.

Borges, G.; Nock, M. K.; Abad, J. M. H.; Hwang, I.; Sampson, N. A.; Alonso, J.; Andrade, L. H.; Angermeyer, M. C.; Beautrais, A.; Bromet, E.; et al. 2010. Twelve-month prevalence of and risk factors for suicide attempts in the world health organization world mental health surveys. *The Journal of clinical psychiatry* 71(12):21777.

Braithwaite, S. R.; Giraud-Carrier, C.; West, J.; Barnes, M. D.; and Hanson, C. L. 2016. Validating machine learning algorithms for twitter data against established measures of suicidality. *JMIR mental health* 3(2):e4822.

Braun, M.; Till, B.; Pirkis, J.; and Niederkrotenthaler, T. 2021. Effects of suicide prevention videos developed by and targeting adolescents: a randomized controlled trial. *European Child & Adolescent Psychiatry* 1–11.

Breckler, S. J. 1984. Empirical validation of affect, behavior, and cognition as distinct components of attitude. *Journal of personality and social psychology* 47(6):1191.

Burnap, P.; Colombo, G.; Amery, R.; Hodorog, A.; and Scourfield, J. 2017. Multi-class machine classification of suicide-related communication on twitter. *Online social networks and media* 2:32–44.

Carmichael, V., and Whitley, R. 2019. Media coverage of robin williams' suicide in the united states: A contributor to contagion? *PloS one* 14(5):e0216543.

Cavanagh, J. T.; Carson, A. J.; Sharpe, M.; and Lawrie, S. M. 2003. Psychological autopsy studies of suicide: a systematic review. *Psychological medicine* 33(3):395–405.

Chung, C., and Pennebaker, J. W. 2007. The psychological functions of function words. *Social communication* 343–359.

Cohn, M. A.; Mehl, M. R.; and Pennebaker, J. W. 2004. Linguistic markers of psychological change surrounding september 11, 2001. *Psychological science*.

Coppersmith, G.; Ngo, K.; Leary, R.; and Wood, A. 2016. Exploratory analysis of social media prior to a suicide attempt. 106–117.

Coppersmith, G.; Leary, R.; Crutchley, P.; and Fine, A. 2018. Natural language processing of social media as screening for suicide risk. *Biomedical informatics insights* 10:1178222618792860.

De Choudhury, M., and Kıcıman, E. 2017. The language of social support in social media and its effect on suicidal ideation risk. In *ICWSM*.

De Choudhury, M.; Gamon, M.; Counts, S.; and Horvitz, E. 2013. Predicting depression via social media.

De Choudhury, M.; Kiciman, E.; Dredze, M.; Coppersmith, G.; and Kumar, M. 2016. Discovering shifts to suicidal ideation from mental health content in social media. 2098–2110.

De Choudhury, M.; Sharma, S.; and Kiciman, E. 2016. Characterizing dietary choices, nutrition, and language in food deserts via social media. In *CSCW*.

Domaradzki, J. 2021. The werther effect, the papageno effect or no effect? a literature review. *International journal of environmental research and public health* 18(5):2396.

Ernala, S. K.; Rizvi, A. F.; Birnbaum, M. L.; Kane, J. M.; and De Choudhury, M. 2017. Linguistic markers indicating therapeutic outcomes of social media disclosures of schizophrenia. *PACM Human-Computer Interaction* (CSCW).

Fahey, R. A.; Matsubayashi, T.; and Ueda, M. 2018. Tracking the werther effect on social media: Emotional responses to prominent suicide deaths on twitter and subsequent increases in suicide. *Social Science & Medicine* 219:19–29.

Fernández-Cabana, M.; Jiménez-Féliz, J.; Alves-Pérez, M. T.; Mateos, R.; Gómez-Reino Rodríguez, I.; and García-Caballero, A. 2015. Linguistic analysis of suicide notes in spain. *The European Journal of Psychiatry* 29(2):145–155.

Fu, K.-w.; Yip, P. S.; et al. 2009. Estimating the risk for suicide following the suicide deaths of 3 asian entertainment celebrities: a meta-analytic approach. *Journal of Clinical Psychiatry* 70(6):869.

Garg, S.; Taylor, J.; El Sherief, M.; Kasson, E.; Aledavood, T.; Riordan, R.; Kaiser, N.; Cavazos-Rehg, P.; and De Choudhury, M. 2021. Detecting risk level in individuals

misusing fentanyl utilizing posts from an online community on reddit. *Internet Interventions* 26:100467.

Guntuku, S. C.; Ramsay, J. R.; Merchant, R. M.; and Ungar, L. H. 2019. Language of adhd in adults on social media. *Journal of attention disorders*.

Handelman, L. D., and Lester, D. 2007. The content of suicide notes from attempters and completers. *Crisis: The Journal of Crisis Intervention and Suicide Prevention* 28(2):102.

Hansen, B. B. 2004. Full matching in an observational study of coaching for the sat. *Journal of the American Statistical Association* 99(467):609–618.

Hawton, K., and van Heeringen, K. 2009. Suicide. *The Lancet* 373(9672):1372–1381.

Holding, T. 1975. Suicide and" the befrienders". *Br Med J* 3(5986):751–752.

Imbens, G. W., and Rubin, D. B. 2015. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Ji, S.; Yu, C. P.; Fung, S.-f.; Pan, S.; and Long, G. 2018. Supervised learning for suicidal ideation detection in online user content. *Complexity* 2018.

Jobes, D. A.; Berman, A. L.; O'Carroll, P. W.; Eastgard, S.; and Knickmeyer, S. 1996. The kurt cobain suicide crisis: perspectives from research, public health, and the news media. *Suicide and Life-Threatening Behavior* 26(3):260–271.

Kher, A.; Johnson, S.; and Griffith, R. 2017. Readability assessment of online patient education material on congestive heart failure. *Advances in preventive medicine* 2017.

Kiciman, E.; Counts, S.; and Gasser, M. 2018. Using longitudinal social media analysis to understand the effects of early college alcohol use.

Kim, K.; Choi, S.; Lee, J.; and Sea, J. 2019. Differences in linguistic and psychological characteristics between suicide notes and diaries. *The Journal of general psychology* 146(4):391–416.

King, K.; Schlichthorst, M.; Reifels, L.; Keogh, L.; Spittal, M. J.; Phelps, A.; and Pirkis, J. 2018. Impacts of a documentary about masculinity and men's health. *American Journal of Men's Health* 12(5):1604–1614.

Kleespies, P. M., and Dettmer, E. L. 2000. An evidence-based approach to evaluating and managing suicidal emergencies. *Journal of Clinical Psychology* 56(9):1109–1130.

Kumar, M.; Dredze, M.; Coppersmith, G.; and De Choudhury, M. 2015. Detecting changes in suicide content manifested in social media following celebrity suicides. In *Proceedings of the 26th ACM conference on Hypertext & Social Media*, 85–94.

Ladwig, K.-H.; Kunrath, S.; Lukaschek, K.; and Baumert, J. 2012. The railway suicide death of a famous german football player: Impact on the subsequent frequency of railway suicide acts in germany. *Journal of affective disorders* 136(1-2):194–198.

Landis, J. R., and Koch, G. G. 1977. The measurement of observer agreement for categorical data. *biometrics* 159–174.

Larson, J. S. 1996. The world health organization's definition of health: Social versus spiritual health. *Social Indicators Research* 38(2):181–192.

Lester, D.; Haines, J.; and Williams, C. L. 2010. Content differences in suicide notes by sex, age, and method: A study of australian suicide notes. *Psychological Reports* 106(2):475–476.

Lester, D. 2004. *Katie's diary: Unlocking the Mystery of a Suicide*. Routledge.

Litvinova, T. A.; Seredin, P. V.; Litvinova, O. A.; and Romanchenko, O. V. 2017. Identification of suicidal tendencies of individuals based on the quantitative analysis of their internet texts. *Computación y Sistemas* 21(2):243–252.

Marchant, A.; Hawton, K.; Stewart, A.; Montgomery, P.; Singaravelu, V.; Lloyd, K.; Purdy, N.; Daine, K.; and John, A. 2017. A systematic review of the relationship between internet use, self-harm and suicidal behaviour in young people: The good, the bad and the unknown. *PloS one* 12(8):e0181722.

Merolli, M.; Gray, K.; and Martin-Sanchez, F. 2013. Health outcomes and related effects of using social media in chronic disease management: a literature review and analysis of affordances. *Journal of biomedical informatics*.

Metzler, H.; Baginski, H.; Niederkrotenthaler, T.; and Garcia, D. 2021. Detecting potentially harmful and protective suicide-related content on twitter: A machine learning approach. *arXiv preprint arXiv:2112.04796*.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 3111–3119.

Morgan, S. L., and Todd, J. J. 2008. 6. a diagnostic routine for the detection of consequential heterogeneity of causal effects. *Sociological Methodology* 38(1):231–282.

Motto, J. A. 1967. Suicide and suggestibility—the role of the press. *American journal of psychiatry* 124(2):252–256.

Motto, J. A. 1970. Newspaper influence on suicide: A controlled study. *Archives of general psychiatry* 23(2):143–148.

Niederkrotenthaler, T., and Till, B. 2020. Effects of awareness material featuring individuals with experience of depression and suicidal thoughts on an audience with depressive symptoms: randomized controlled trial. *Journal of behavior therapy and experimental psychiatry* 66:101515.

Niederkrotenthaler, T.; Till, B.; Kapusta, N. D.; Voracek, M.; Dervic, K.; and Sonneck, G. 2009. Copycat effects after media reports on suicide: A population-based ecologic study. *Social science & medicine* 69(7):1085–1090.

Niederkrotenthaler, T.; Voracek, M.; Herberth, A.; Till, B.; Strauss, M.; Etzersdorfer, E.; Eisenwort, B.; and Sonneck, G. 2010. Role of media reports in completed and

prevented suicide: Werther v. papageno effects. *The British Journal of Psychiatry* 197(3):234–243.

Niederkrotenthaler, T.; Braun, M.; Pirkis, J.; Till, B.; Stack, S.; Sinyor, M.; Tran, U. S.; Voracek, M.; Cheng, Q.; Arendt, F.; et al. 2020. Association between suicide reporting in the media and suicide: systematic review and meta-analysis. *Bmj* 368.

Niederkrotenthaler, T.; Till, B.; Kirchner, S.; Sinyor, M.; Braun, M.; Pirkis, J.; Tran, U. S.; Voracek, M.; Arendt, F.; Ftanou, M.; et al. 2022. Effects of media stories of hope and recovery on suicidal ideation and help-seeking attitudes and intentions: systematic review and meta-analysis. *The Lancet Public Health* 7(2):e156–e168.

Nobles, A. L.; Glenn, J. J.; Kowsari, K.; Teachman, B. A.; and Barnes, L. E. 2018. Identification of imminent suicide risk among young adults using text messages. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–11.

Nock, M. K., and Kazdin, A. E. 2002. Examination of affective, cognitive, and behavioral factors and suicide-related outcomes in children and young adolescents. *Journal of clinical child and adolescent psychology* 31(1):48–58.

O'Connor, R. C., and Nock, M. K. 2014. The psychology of suicidal behaviour. *The Lancet Psychiatry* 1(1):73–85.

O'dea, B.; Wan, S.; Batterham, P. J.; Calear, A. L.; Paris, C.; and Christensen, H. 2015. Detecting suicidality on twitter. *Internet Interventions* 2(2):183–188.

Organization, W. H., et al. 2014. *Preventing suicide: A global imperative*. World Health Organization.

Pennebaker, J. W., and Chung, C. K. 2007. Expressive writing, emotional upheavals, and health. *Handbook of health psychology* 263–284.

Pennebaker, J. W., and Seagal, J. D. 1999. Forming a story: The health benefits of narrative. *Journal of clinical psychology* 55(10):1243–1254.

Pennebaker, J. W.; Francis, M. E.; and Booth, R. J. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates* 71(2001):2001.

Pennebaker, J. W.; Mehl, M. R.; and Niederhoffer, K. G. 2003. Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology* 54(1):547–577.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Phillips, D. P. 1974. The influence of suggestion on suicide: Substantive and theoretical implications of the werther effect. *American sociological review* 340–354.

Phillips, D. P. 1978. Airplane accident fatalities increase just after newspaper stories about murder and suicide. *Science* 201(4357):748–750.

Reynders, A.; Kerkhof, A. J.; Molenberghs, G.; and Van Audenhove, C. 2015. Help-seeking, stigma and attitudes of people with and without a suicidal past. a comparison between a low and a high suicide rate country. *Journal of Affective Disorders* 178:5–11.

Robins, J. M.; Hernan, M. A.; and Brumback, B. 2000. Marginal structural models and causal inference in epidemiology.

Romer, D.; Jamieson, P. E.; and Jamieson, K. H. 2006. Are news reports of suicide contagious? a stringent test in six us cities. *Journal of Communication* 56(2):253–270.

Rosenbaum, P. R., and Rubin, D. B. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70(1):41–55.

Rosenbaum, P. R., and Rubin, D. B. 1984. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American statistical Association* 79(387):516–524.

Rosenblat, J. D.; Kakar, R.; and McIntyre, R. S. 2016. The cognitive effects of antidepressants in major depressive disorder: a systematic review and meta-analysis of randomized clinical trials. *International Journal of Neuropsychopharmacology* 19(2).

Rozé, J.-C.; Cambonie, G.; Marchand-Martin, L.; Gournay, V.; Durrmeyer, X.; Durox, M.; Storme, L.; Porcher, R.; Ancel, P.-Y.; Group, H. E. . S.; et al. 2015. Association between early screening for patent ductus arteriosus and in-hospital mortality among extremely preterm infants. *Jama* 313(24):2441–2448.

Rubin, D. B. 2005. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association* 100(469):322–331.

Rude, S.; Gortner, E.-M.; and Pennebaker, J. 2004. Language use of depressed and depression-vulnerable college students. *Cognition & Emotion* 18(8):1121–1133.

Saha, K., and Sharma, A. 2020. Causal factors of effective psychosocial outcomes in online mental health communities. In *ICWSM*.

Saha, K.; Sugar, B.; Torous, J.; Abrahao, B.; Kıcıman, E.; and De Choudhury, M. 2019. A social media study on the effects of psychiatric medication use. 13:440–451.

Saha, K.; Torous, J.; Caine, E. D.; De Choudhury, M.; et al. 2020. Psychosocial effects of the covid-19 pandemic: large-scale quasi-experimental study on social media. *Journal of medical internet research* 22(11):e22600.

Saha, K.; Liu, Y.; Vincent, N.; Chowdhury, F. A.; Neves, L.; Shah, N.; and Bos, M. W. 2021a. Advertiming matters: Examining user ad consumption for effective ad allocations on social media. In *Proc. CHI*.

Saha, K.; Seybolt, J.; Mattingly, S. M.; Aledavood, T.; Konjeti, C.; Martinez, G. J.; Grover, T.; Mark, G.; and De Choudhury, M. 2021b. What life events are disclosed on social media, how, when, and by whom? In *Proceedings of the 2021 CHI conference on human factors in computing systems*, 1–22.

Saha, K.; Weber, I.; and De Choudhury, M. 2018a. A social media based examination of the effects of counseling recommendations after student deaths on college campuses. In *Twelfth International AAAI Conference on Web and Social Media*.

Saha, K.; Weber, I.; and De Choudhury, M. 2018b. A social media based examination of the effects of counseling recommendations after student deaths on college campuses. In *ICWSM*.

Sawhney, R.; Manchanda, P.; Singh, R.; and Aggarwal, S. 2018. A computational approach to feature extraction for identification of suicidal ideation in tweets. In *Proceedings of ACL 2018, Student Research Workshop*, 91–98.

Schafer, J. L., and Kang, J. 2008. Average causal effects from nonrandomized studies: a practical guide and simulated example. *Psychological methods* 13(4):279.

Schaffer, J. M.; Singh, S. K.; Reitz, B. A.; Zamanian, R. T.; and Mallidi, H. R. 2015. Single-vs double-lung transplantation in patients with chronic obstructive pulmonary disease and idiopathic pulmonary fibrosis since the implementation of lung allocation based on medical need. *Jama* 313(9):936–948.

Scott, K. M.; Hwang, I.; Chiu, W.-T.; Kessler, R. C.; Sampson, N. A.; Angermeyer, M.; Beautrais, A.; Borges, G.; Bruffaerts, R.; De Graaf, R.; et al. 2010. Chronic physical conditions and their association with first onset of suicidal behavior in the world mental health surveys. *Psychosomatic Medicine* 72(7):712–719.

Setoguchi, S.; Schneeweiss, S.; Brookhart, M. A.; Glynn, R. J.; and Cook, E. F. 2008. Evaluating uses of data mining techniques in propensity score estimation: a simulation study. *Pharmacoepidemiology and drug safety* 17(6):546–555.

Shing, H.-C.; Nair, S.; Zirikly, A.; Friedenberg, M.; Daumé III, H.; and Resnik, P. 2018. Expert, crowdsourced, and machine assessment of suicide risk via online postings. In *Proceedings of the fifth workshop on computational linguistics and clinical psychology: from keyboard to clinic*, 25–36.

Song, J.; Song, T. M.; Seo, D.-C.; and Jin, J. H. 2016. Data mining of web-based documents on social networking sites that included suicide-related words among korean adolescents. *Journal of Adolescent Health* 59(6):668–673.

Stack, S. 1983. The effect of the jonestown suicides on american suicide rates. *The Journal of social psychology* 119(1):145–146.

Stack, S. 1987. Celebrities and suicide: A taxonomy and analysis, 1948-1983. *American sociological review* 401–412.

Stack, S. 1990. A reanalysis of the impact of non celebrity suicides. *Social psychiatry and psychiatric epidemiology* 25(5):269–273.

Staffa, S. J., and Zurakowski, D. 2018. Five steps to successfully implement and evaluate propensity score matching in clinical research studies. *Anesthesia & Analgesia* 127(4):1066–1073.

Steinke, J.; Root-Bowman, M.; Estabrook, S.; Levine, D. S.; and Kantor, L. M. 2017. Meeting the needs of sexual and gender minority youth: formative research on potential digital health interventions. *Journal of Adolescent Health* 60(5):541–548.

Stirman, S. W., and Pennebaker, J. W. 2001. Word use in the poetry of suicidal and nonsuicidal poets. *Psychosomatic medicine* 63(4):517–522.

Stuart, E. A. 2010. Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics* 25(1):1.

Tadesse, M. M.; Lin, H.; Xu, B.; and Yang, L. 2020. Detection of suicide ideation in social media forums using deep learning. *Algorithms* 13(1):7.

Till, B.; Strauss, M.; Sonneck, G.; and Niederkrotenthaler, T. 2015. Determining the effects of films with suicidal content: a laboratory experiment. *The British Journal of Psychiatry* 207(1):72–78.

Till, B.; Tran, U. S.; Voracek, M.; and Niederkrotenthaler, T. 2017. Beneficial and harmful effects of educative suicide prevention websites: randomised controlled trial exploring papageno v. werther effects. *The British Journal of Psychiatry* 211(2):109–115.

Till, B.; Arendt, F.; Scherr, S.; and Niederkrotenthaler, T. 2018. Effect of educative suicide prevention news articles featuring experts with vs without personal experience of suicidal ideation: a randomized controlled trial of the papageno effect. *The Journal of clinical psychiatry* 80(1):13780.

Tousignant, M.; Mishara, B. L.; Caillaud, A.; Fortin, V.; and St-Laurent, D. 2005. The impact of media coverage of the suicide of a well-known quebec reporter: the case of gaetan girouard. *Social science & medicine* 60(9):1919–1926.

VandenBos, G. R. 2007. *APA dictionary of psychology.* American Psychological Association.

Verma, G.; Bhardwaj, A.; Aledavood, T.; De Choudhury, M.; and Kumar, S. 2022. Examining the impact of sharing covid-19 misinformation online on mental health. *Scientific Reports* 12(1):8045.

Wang, Q.; Saha, K.; Gregori, E.; Joyner, D.; and Goel, A. K. 2021. Mutual theory of mind in human-ai interaction: How language reflects what students perceive about a virtual teaching assistant. In *Proc. CHI.*

Wasserman, I. M. 1984. Imitation and suicide: A reexamination of the werther effect. *American sociological review* 427–436.

WHO. 2019. Suicide in the world: global health estimates.

Xiang, X.; Lu, X.; Halavanau, A.; Xue, J.; Sun, Y.; Lai, P. H. L.; and Wu, Z. 2021. Modern senicide in the face of a pandemic: An examination of public discourse and sentiment about older adults and covid-19 using machine learning. *The Journals of Gerontology: Series B* 76(4):e190–e200.

Yao, L.; Chu, Z.; Li, S.; Li, Y.; Gao, J.; and Zhang, A. 2021. A survey on causal inference. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 15(5):1–46.

Yeasmin, N.; Mahbub, N. I.; Baowaly, M. K.; Singh, B. C.; Alom, Z.; Aung, Z.; and Azim, M. A. 2022. Analysis and prediction of user sentiment on covid-19 pandemic using tweets. *Big Data and Cognitive Computing* 6(2):65.

Yuan, Y.; Verma, G.; Keller, B.; and Aledavood, T. 2022. The impact of covid-19 pandemic on lgbtq online communities. *arXiv preprint arXiv:2205.09511*.