IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 9 May 2023, accepted 11 June 2023, date of publication 26 June 2023, date of current version 3 July 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3289295

RESEARCH ARTICLE

Forecasting National-Level Self-Harm Trends With Social Networks

SUPPAWONG TUAROB[1], (Member, IEEE), KRITTIN CHATRINAN[1], THANAPON NORASET[1], TANISA TAWICHSRI[2], AND TIPAJIN THAIPISUTIKUL[1]

[1]Faculty of Information and Communication Technology, Mahidol University, Salaya 73170, Thailand
[2]Puey Ungphakorn Institute for Economic Research, Bank of Thailand, Bangkok 10200, Thailand

Corresponding author: Suppawong Tuarob (suppawong.tua@mahidol.edu)

This research project is supported by Mahidol University (Fundamental Fund: fiscal year 2023 by National Science Research and Innovation Fund (NSRF)).

ABSTRACT Self-harm pertains to actions of self-inflicted poisoning or injury that lead to either non-fatal injuries or death, irrespective of the individual's intention. Self-harm incidents not only cause loss to individuals but also incur a negative impact on the nation's economy. Studies have demonstrated an increase in trends of self-harm that are correlated with the emergence of technological advancements and swift urban expansion in developing countries. The capacity to nowcast and forecast national-level patterns of self-harm trends could be imperative to policymakers and stakeholders in the public health sector, as it would enable them to implement prompt measures to counteract the underlying factors or avert these projected calamities. Prior research has utilized historical data to predict self-harm trends at the population level in various nations using conventional statistical forecasting methods. However, in some countries, such historical statistics may be challenging to obtain or insufficient for accurate prediction, impeding the ability to comprehend and project the national self-harm landscape in a timely manner. This paper proposes *FAST*, a framework designed to forecast self-harm patterns at the national level by analyzing mental signals obtained from a large volume of social media data. These signals serve as a proxy for real-world population mental health that could be used to enhance the forecastability of self-harm trends. Specifically, language-agnostic language models are first trained to extract different mental signals from collected social media messages. Then, these signals are aggregated and processed into multi-variate time series, on which the time-delay embedding algorithm is applied to transform into temporal embedded instances. Finally, various machine learning regressors are validated for their forecastability. The proposed method is validated through a case study in Thailand, which utilizes a set of 12 mental signals extracted from tweets to forecast death and injury cases resulting from self-harm. The results show that the proposed method outperformed the traditional ARIMA baseline by 43.56% and 36.48% on average in terms of MAPE on forecasting death and injury cases from self-harm, respectively. As far as current understanding permits, our research represents the initial exploration of utilizing aggregated social media information for the purposes of nowcasting and forecasting trends of self-harm on a nationwide scale. The results not only provide insight into improved forecasting techniques for self-harm trends but also establish a foundation for forthcoming social-network-driven applications that hinge on the capacity to predict socioeconomic factors.

INDEX TERMS Self-harm, nowcasting, forecasting, online social networks, cross-lingual text classification.

I. INTRODUCTION

Self-harm refers to intentional self-poisoning or self-injury, regardless of the nature of the motivation or the severity of suicidal intent [35], that could result in injury or death. Self-harm and suicide have been prevalent problems, especially in developing countries [9]. According to a recent study, a significant proportion of suicide cases, approximately 77%, were observed in low- and middle-income countries [47]. This trend has been associated with the uptake

The associate editor coordinating the review of this manuscript and approving it for publication was M. Shamim Kaiser.

of technological advancements and the rapid pace of urbanization in these regions [56]. The exacerbation of incidents involving self-harm not only results in personal grief and loss but also has enduring adverse effects on the economy, primarily due to the reduction in long-term labor productivity [41]. The ability to monitor and forecast population-level self-harm trends could prove vital to national-level policymakers and public healthcare stakeholders in devising means to timely gauge the situations and implement procedures to neutralize or prevent such anticipated tragedies [70]. For example, upon being informed that certain stringent policies aimed at addressing nationwide epidemics have led to mental health issues among citizens and are expected to contribute to a significant increase in self-harm trends, policymakers may consider making appropriate modifications to the current policies that are causing these issues. Furthermore, the implementation of public health interventions, such as mobile psychiatry units or hotlines, could be deployed to target populations experiencing adverse effects. Presently, the techniques employed to acquire knowledge about self-harm trends at the national level depend on administrative reports from healthcare centers and hospitals across the country. This approach necessitates significant financial, human, and time resources, leading to infrequent and delayed data availability. Statistics that are coarse-grained and delayed may have limited utility for the purpose of proactive policy-making.

The need for monitoring and forecasting trends in self-harm and suicide has been highlighted in recent literature [63]. The problem has been framed as a time series forecasting task, with conventional techniques such as ARIMA and the Holt-Winters methods being applied [71], [90]. Recent research has illuminated the diverse individual and external factors that influence the decision-making process of individuals to engage in acts of self-harm [22], [66]. According to Chang and Lee [17], relying solely on historical self-harm and suicide statistics may not be adequate for developing precise forecasting models. This highlights the need for supplementary data sources that capture the population's reactions to current self-harm and suicide trends. In this direction, previous literature has examined the potential of utilizing Google Trends data, specifically search queries related to self-harm and suicide, to enhance the accuracy of national-level self-harm forecasting [39]. However, recent research has indicated that Google Trends may not be an appropriate proxy for measuring aggregate self-harm behavior. This is mainly due to the undisclosed algorithm that governs the mechanism of Google Trends [25], as well as the assumption that there is a great overlap between the population that generates Google search queries and those who actually engage in self-harm acts [91]. In addition, a recent academic investigation has revealed that the employment of self-harm-related keywords in Google Trends statistics did not correlate strongly with actual self-harm statistics in Thailand [62].

To fill the gap in the lack of a good timely indicator, we propose utilizing extensive online social media data as a feasible proxy for collective self-harm behavior. The potential connections between social media platforms, including Twitter and Facebook, and suicidal ideation have been explored in previous research studies [53], [82], [84]. Conversely, other studies have employed social media to monitor and forecast incidents of self-harm and suicide among individuals [55], [78]. Nevertheless, the utilization of social media as a means of predicting self-harm and suicide tendencies at an aggregate level has yet to be explored.

According to research conducted by Wang et al. [95], individuals with mental disorders tend to feel more at ease expressing their thoughts on social media platforms. Not only that, the usage of social media has been linked to worsened mental health, especially among young adults, [12], as well as higher rates of suicide and self-harm [1]. Therefore, it is reasonable to infer that indicators extracted from social media could be a good proxy for the rising rates of population-level self-harm, since a significant portion of social media users may either openly discuss negative mental conditions on these platforms or their negative mental health states can be implicitly inferred from their expressions on social media. Under this conjecture, the characteristics that signify the cognitive states of the message's author may function as a proxy for their current or future mental health states and behaviors. According to a recent study [62], there exists a strong correlation between collective mental signals, including fear, sadness, disgust, and suicidal tendencies, extracted from social media messages and the national incidence of injuries and deaths resulting from self-harm in Thailand. Building upon prior research, our proposed framework involves utilizing extensive social media data as a proxy for measuring the population's mental health, which can be leveraged to forecast nationwide self-harm patterns.

Specifically, we propose *FAST*, a novel framework for *F*orecasting *A*ggregate-level *S*elf-harm *T*rends using large-scale social media data, as part of the PSIMILAN (A data processing and visualization system for *PS*ychological *I*mpact on *M*ental health us*I*ng *LA*rge-scale social *N*etworks) project. The framework first employs relevant keywords to gather publicly available social media data. The process involves extracting mental signals such as sentiments, emotions, and suicidal tendencies from individual messages, which are then aggregated to form multivariate time series at a national level. These temporal mental signals are then combined with historical self-harm statistics to generate machine-learning-based forecasting models that predict the number of injuries and deaths from reported self-harm incidents at various forecasting horizons. The proposed framework exhibits generalizability across diverse regional contexts and social media platforms, attributed to the utilization of language-agnostic language models for mental signal extraction, which solely rely on the existence of textual content in social media data regardless of composing languages.

A case study examining mental signals derived from social media data to predict self-harm trends in Thailand is used to validate the proposed method. The findings indicate that the optimal configuration of the proposed forecasting models produces an average mean absolute percentage error (MAPE) of 13.27% and 10.15% for forecasting death and injury cases resulting from self-harm, respectively, across the horizons of zero to six months. These results surpass the ARIMA baseline by 43.56% and 36.48%, respectively. The proposed method exhibits optimal nowcasting performance for death and injury cases at horizon = 0 (nowcasting), with MAPE of 6.59% and 6.20%, respectively. However, it is observed that the average errors tend to escalate as the forecasting horizon increases. With the ability to accurately forecast self-harm trends at the national level, it is our overarching objective for the proposed framework to be implemented as part of a decision support system capable of visualizing and investigating the impact of policies or situations on the population's self-harm behavior at both the national and regional levels. The implementation of such a system has the potential to aid policymakers in making well-informed and proactive decisions regarding the development or modification of policies that could have an impact on the mental health of individuals.

Concretely, this paper presents the following key contributions:

- We establish social media data as a viable additional data source to improve the forecasting of population-level self-harm trends.
- We propose *FAST*, a framework for developing self-harm forecasting models at the population level using collective mental signals extracted from large-scale social media data. The framework is composed of several interdependent modules that collaborate to obtain data from social media, extract and consolidate mental signals, incorporate time-delay information, select appropriate machine-learning regression models, and systematically evaluate various model configurations.
- We conduct thorough empirical evaluations using the standard step-wise multi-variate time series cross-validation evaluation protocol. Our study utilizes the traditional ARIMA model as the baseline and explores various configurations of target attributes, regression models, lags, and horizons.
- We make the data and source code available for research purposes at https://github.com/krittintey/psimilan-fast.

The rest of this paper is organized as follows. Section II reviews literature related to self-harm and suicide prediction and forecasting both at the individual and aggregate levels. Section III presents our proposed framework in detail. Section IV discusses the datasets used as well as the key experiment results. Should the proposed framework be adopted for real-world applications, ethical and societal ramifications are expected to arise, some of which are briefly examined in Section V. Finally, Section VI concludes the paper.

II. BACKGROUND AND RELATED WORK

The ability forecast the aggregate-level self-harm trends could prove useful not only to policymakers in analyzing the root causes for policy interventions but also to public healthcare stakeholders for dispatching remedies to mitigate and combat the issues [45]. Indeed, the ability to forecast self-harm and suicide rates has been investigated since 1978, where simple statistical techniques, such as the Box-Jenkins method, were used for forecasting suicidal rates using historical statistics [94]. The advent of Internet technologies has facilitated convenient access to online information, most of which through a web search. Later studies have investigated utilizing these web search queries as proxies to monitor suicide trends at the national level [10], [91]. However, in certain developing countries, for example, Thailand, these web search queries were not found to correlate to self-harm and suicide trends, possibly due to the underlying population which produced these search queries, which could be different from those actually committing self-harm [62]. Recently, with the rise of online social networking services as alternative and complement platforms for communication, several studies have investigated both their contribution to causes of self-harm and suicide [53], [59] and purposes as a proxy to assess self-harm and suicide risks in potential mental disorder patients [28], [69]. Online social networking data has been established as potential sources for monitoring real-world phenomena both at the individual and aggregate levels. For example, Twitter and Facebook data have been used to predict individual flu infection [73] and forecast the trends at the national level [67]. Many studies have extended the use of social media data for tracing and forecasting self-harm and suicide risks in individuals [11], [40]; however, the use of such large-scale user-generated data for forecasting population-level self-harm and suicide trends has yet to surface. Therefore, to provide a comprehensive background and related work, this section first discusses previously proposed methods for individual-level self-harm prediction. Then, the next subsection reviews related work on forecasting self-harm and suicide trends at the national level, both using traditional historical statistics and online information (e.g., Google Trends).

A. INDIVIDUAL-LEVEL SELF-HARM PREDICTION

The ability to trace and recognize self-harm risks in individuals could prove critical in preventing subsequent tragedies and losses. The fact that many people who commit suicide come in contact with healthcare systems prior to their death [40] enables collecting necessary information, such as patients' background, health, behaviors, and social engagement, for tracking and assessing individual self-harm risks. Edgcomb et al. [21] proposed a machine-learning approach for predicting suicidal and self-harm behavior of patients after hospitalization. Their approach trains the CART model [49] with information from the electronic health records (EHRs), such as gender, race, age, psychiatric diagnoses, substance use disorder, and medications, to predict whether a patient will be hospitalized for suicidal or self-harm attempts within 365 days after being last discharged. Kyron et al. [44] presented a study that investigated the use of the hierarchical logistic regression model to predict self-harm incidents among patients within inpatient psychiatric facilities. The classifier was trained with daily self-reports from test subjects and was used to predict individual self-harm risks. They found that the increases in a wish to die and psychological distress were directly associated with the increased risk of self-harm. They further suggested that the findings could be extended to implementing monitoring systems for identifying short-term fluctuations in mood and interpersonal circumstances related to self-harm risks. Recently, Liem et al. [47] investigated the possibility of predicting individual self-harm and suicide ideation during the COVID-19 pandemic from information collected as part of the nationwide survey. The questionnaire comprises information pertaining to demography, loneliness from social isolating, and self-harm and suicide ideation assessment. A case study of 5,211 participants in Indonesia who conducted the survey from May 25 - June 16, 2021, was used to validate the efficacy of the hierarchical logistic regression model. The findings revealed that highly important features were age, residence, job, religion, gender, sexual orientation, HIV status, disability, and loneliness.

Besides using the patient's health records and self-administered questionnaire information to assess the risks of self-harm and suicide, recent studies have shed light on the emergence of online social networks and their relations with individual self-harm and suicide. While some studies have identified that certain social network activities could increase self-harm risks, especially for those with pre-existing mental conditions [53], others have utilized these social media data as a proxy to comprehend thoughts of potential self-harm users. Indeed, related work that utilizes social media data for individual-level self-harm analysis can be divided into two groups. The first refers to investigating and developing techniques for discovering self-harm-related content in a massive pool of social media data, often framing the problem as a text classification problem. The latter focuses on identifying social media users who have high risks of self-harm, where the users' corresponding social media posts are automatically digested and analyzed to quantify the self-harm risk.

The sheer amount of social media data generated daily is enormous, encompassing diverse topics and purposes generated by a wide variety of users [29]. The ability to automatically identify and extract a relevant subset of this large-scale user-generated data has been proved useful in monitoring, analytics, and prediction applications in many domains such as healthcare [3], finance [42], politics [6], and public policies [87]. The presence of self-harm content in social networking services has recently raised awareness of the research communities to develop automated intelligent techniques to address multiple angles of the problems. As evidence, the Conference and Labs of the Evaluation Forum (CLEF) has organized the eRisk task "Early Detection of Signs of Self-Harm" during 2019 - 2021 [51], [52], [65] to provide evaluation benchmarks for developing methods to detect self-harm-indicating messages in the Reddit platform.

In computational psycholinguistics, Ríssola et al. [77] found that messages composed by users with abnormal mental health could exhibit certain linguistic patterns, such as the use of adverbs, verb tenses, and topic-specific vocabulary. Furthermore, they found that users with mental disorders tend to expose their emotions more regularly than control individuals. With such knowledge, it would be possible to build automated routines to capture characteristics in language styles used by potential self-harm users. Often, this type of work assumes that social media data comprises a set of communication messages and frames the problem as a text classification task [5], [75]. Wang et al. [95] was among the first research teams to explore this dimension. They discovered that self-harm users found it less difficult to communicate self-harm-related thoughts and behaviors on social media than in person. Evaluating the dataset collected from Flickr reveals that their proposed Self-harm Content Prediction (SCP) method outperformed the traditional word-embedding and CNN-based methods in the self-harm message detection task. Later, several classical and deep learning approaches were adopted and developed for detecting self-harm and suicide content in online social networks [55].

The ability to surveil the online realm for the presence of self-harm content is not only crucial for the monitoring tasks but also serves as a building block for identifying social media users who have high risks of self-harm. Roy et al. [81] proposed to do so by first quantifying different mental signals from a tweet using neural networks, including stress, loneliness, burdensomeness, hopelessness, depression, anxiety, and insomnia. Then, these extracted signals were used to train random forest models to detect suicidal ideation events that can be used to assess suicide risks at the user level. A case study of 283 Twitter users with suicidal ideation and 2,655 control users was used to validate their proposed Suicide Artificial Intelligence Prediction Heuristic (SAIPH) approach. Later, for the CLPsych 2021 Shared Task, Gollapalli et al. [32] developed the Self-Harm Topic Model (SHTM) for identifying those at risk for suicide based on their tweets. Their method combines Latent Dirichlet Allocation (LDA) and a self-harm dictionary for modeling target users' tweets. Then, features based on self-harm mood changes and topic changes in tweets over time were used to train a deep-learning model to predict suicidal users. Recently, Cao et al. [15] devised and implemented a personal suicide-oriented knowledge graph for detecting suicidal intent on social media. An attention mechanism with two network layers was used to explain and identify important risk variables for suicidal thoughts in social media users. The findings of the study conducted on 7,329 Sina Weibo microblog users, comprising 3,652 individuals with suicidal ideation and 3,677 individuals without suicidal ideation, yielded noteworthy outcomes. The algorithm demonstrated a high level of accuracy, exceeding that of other contemporary methods, by utilizing personal knowledge graph information in identifying users with suicidal ideation, achieving an accuracy of over 93%. Moreover, within the six classifications of individual factors, the three foremost crucial indicators were occupation, disposition, and expertise. The detection of suicidal ideation is primarily influenced by various factors such as the content of the posted text, the level and duration of stress, the presence of posted pictures, and the occurrence of ruminative thinking.

The above-reviewed studies utilized various sources of personal information from available health records, self-administered questionnaires, and social media information for predicting self-harm risks at either the message or individual level. While several studies showed promising evaluation results on their corresponding datasets, proposed approaches in this direction have faced limitations in terms of applicability. First, despite the obvious benefit of predictive technologies for self-harm assessment in individuals, implementing and adopting these tools in clinical settings still yield little value or, in some cases, ''might do more harm than good'' [40]. Second, for aggregate-level applications such as analyzing and crafting policies or public health strategies to combat rising national-level self-harm problems, policymakers would benefit more from decision support systems that accurately predict and forecast the aggregate trends of self-harm at the national level rather than the individual level.

B. AGGREGATE-LEVEL SELF-HARM PREDICTION

Societal loss from self-harm and suicide has become a global concern [31], which is constantly growing, mainly owing to the rapid rise in technology adoption [93] and urbanization [4]. A traditional way to obtain aggregate-level self-harm statistics is by collecting reports from hospitals and healthcare centers in the country. However, such a method typically involves manual data processing and corporation from nationwide healthcare units, which often results in delays for several months. Such delay in data availability could hinder the policymakers' ability to effectively handle ongoing or unforeseen national problems. Hence, the capacity to precisely assess and predict the general patterns of self-harm incidents on a national scale may be advantageous for policymakers and public health stakeholders in formulating prompt strategies to address such circumstances. The scientific literature has put forth various methodologies for predicting trends in self-harm behavior through the analysis of historical data and the application of traditional statistical techniques. Chang and Lee [17] raised an observation that studies in suicide rate forecasting had been limited primarily due to various complex factors that affected suicidal behaviors, which resulted in unsatisfactory predictions. They then used the experience curve to forecast annual suicide rates and numbers in 15 countries in 2010, 2020, and 2030. Preti and Lentini [71] proposed to use a combination of statistical methods such as ARIMA, the Holt-Winters seasonal method, the ETS model, and the TBATS model to forecast monthly numbers of suicide cases in Italy. Separate forecasting models were validated for male and female suicide cases, where they found that the forecastability for male suicide cases was clearer than that of females. Rostami et al. [80] used an exponential smoothing state-space model to forecast monthly suicidal rates in the West of Iran. They concluded that there were no significant observable trends of suicide over the study period of March 2006 to September 2013. However, they were able to detect statistically significant variations that called for additional data from other parts of the country with longer duration to better estimate the general suicide trend at the country level. Recently, Swain et al. [90] also pointed out that the suicide rates in India have been increasing, aligning with the global trend. They experimented with various statistical forecasting techniques on the historical suicide rates collected from India's National Crime Record Bureau Reports during 1969 - 2018 and found that ARIMA was the most effective in forecasting suicide rates in India for the next decade.

Several studies have investigated the risk factors that influence self-harm ideation and found that the risk factors varied across different groups of people depending on gender, age ranges, careers, and geographical regions [16], [22], [24], [33]. Some of these risk factors include substance misuse, poor family, and peer relationships [74]. While most of these risk factors depend on individual circumstances, recent studies revealed that some factors could also be affected by ongoing societal situations such as financial status and terrifying disease infection [36], [66]. With such diverse influencing factors that govern the decision to commit self-harm, there is no wonder why it is difficult for prediction models that learn only from the historical time series alone to yield satisfactory forecastability [17]. Indeed, recent studies have explored the use of auxiliary aggregate-level information that could reflect the population's opinions, such as web search queries (e.g., Google Trends) which could be used to represent the collective needs of particular information at different temporal and thematic scales [38]. Such prevalent need-to-know information pertaining to certain topics could reveal the current ongoing phenomenon in a society that potentially serves as latent variables in aggregate-level prediction tasks. Barros et al. [10] found that incorporating the search volumes of queries related to depression and suicide from Google Trends into a neural network autoregression model helped to improve the forecasting of suicide rates in Ireland with MAE ranging between 4.14 and 9.61. Following prior research findings that absolutist thinking could be a sign of depression and self-harm, Adam-Troian and Arciszewski [2] explored the use of absolutist keywords (e.g., ''completely'' and ''totally'') to query the corresponding Google Trends time series and incorporated them into a mixed model to forecast state-level suicide rates in the United States. Their experiment results evidenced the link between the absolutist linguistic markers and suicide deaths at the collective level. Recently, Kandula et al. [39] proposed a two-stage model for forecasting state-level suicide mortality, utilizing search queries from Google Trends queried by suicide-related terms and call logs from the crisis hotline services. The first stage uses ARIMA to produce four forecasting models separately trained with historical suicide mortality, search queries, call logs, and the combination of search queries and call logs. The outputs of these four models are then used as inputs for the second-stage ARIMA model to forecast suicide mortality in the next six months in 50 states in the United States.

While Google Trends has been used for forecasting self-harm trends, such a user-generated information source may not be suitable as a proxy to learn the underlying population-level factors that influence national self-harm trends due to the following reasons. First, a recent discovery has presented criticisms about using Google Trends for research purposes [25], largely because the algorithm that governs the behavior of such a service has not been published, resulting in unexplainable phenomena such as trend changes even when queried with the same keywords and time frames. Furthermore, specifically for the problem at hand, Google Trends has been reported for low validity in forecasting national suicide rates [91]. The main reasons stem from a number of assumptions that have not been statistically proven valid: 1) the underlying population that generates Google search queries comprises mostly those that actually have self-harm ideation, and 2) web-search behavior is strongly linked with suicide behavior. Indeed, recent findings [62] revealed a weak correlation between suicide-related Google Trends volumes and the actual national self-harm statistics in Thailand.

Online social networks have emerged as a means of communication among online users and serve as user-generated information sources alternative to web search queries. In contrast to web search users who primarily seek information, social media users interact with each other. As a result, social media has been established as a viable source for mining public opinions that become proxies for many applications that aim to assess real-world events [57]. While social media users have been questioned for their ability to represent the whole population [58], studies have found that information extracted from large-scale social media data exhibited strong correlations with real-world phenomena at the aggregate level. In computational psychology, studies have found a strong tie between self-harm behavior and social media usage, especially in high-income countries where digital technologies and literacy have fully propagated [64]. However, analyses of the relationship between self-harm ideation and collective social media engagement in developing countries still fall short. Thus, this research proposes the utilization of extensive online social networks as an alternative indicator for predicting self-harm trends at the national level. To our knowledge, this research endeavors to investigate the feasibility of utilizing knowledge obtained from collective social media data to forecast trends of self-harm cases. Closest to addressing this research problem is the work by Noraset et al. [62], who discovered a strong correlation between mental signals, including fear, sadness, disgust, and suicidal tendencies, extracted from aggregated social media data and the real-world incidence of self-harm injuries and fatalities at the national level in Thailand. Based on their findings, we propose a framework for forecasting self-inflicted injuries and fatalities at the national level, utilizing mental cues derived from extensive user-generated social media data. The outcomes of this investigation have the potential to facilitate the development of decision-support systems for policymakers and public healthcare practitioners. These systems could aid in the creation and execution of strategies to address anticipated rises in self-harm trends. Additionally, they could enable the monitoring of the effects of implemented policies on national self-harm behaviors in a timely and cost-efficient manner.

III. METHODOLOGY

While several studies have defined self-harm concepts differently and even used the terms self-harm and suicide interchangeably [50], this research follows the definition of self-harm by Hawton et al. [35]. Specifically, regardless of the nature of the motivation or the level of suicidal intent, self-harm is referred to as deliberate self-poisoning or self-injury that may result in injury or death. Therefore, with such a definition, attempted suicide, successful or unsuccessful, is also considered a self-harm act. This definition of self-harm is used here rather than the binary classification of such actions as non-suicidal self-injury and attempted suicide as early used in the literature [34] because suicidal intent is a multidimensional phenomenon where the patient's and clinician's perceptions of suicidal intent might disagree [35]. Furthermore, national clinical guidelines focus on self-harm for implementing relevant management strategies [14], [46].

The ability to forecast population-level self-harm trends could therefore allow public health practitioners and policymakers to anticipate events that could trigger self-harm risk factors and invent intermediate strategies for effective self-harm management in a timely manner. The main novelty of this research is the use of mental signals extracted from large-scale social media data as a proxy to forecast aggregate self-harm trends. The motivation to investigate the viability of such user-generated data stems from prior research on the risk factors that influence individual decisions to commit self-harm, most of which result from mental health disorders either caused by existing psychiatric conditions or external factors such as being bullied [37] or strict governmental policies in combating grave situations [96]. Recent studies suggested that the increasing self-harm trends are linked to the expansion of social media adoption [59], and people with self-harm ideation found themselves more comfortable expressing their thoughts on social networks [95]. Drawing upon this knowledge, the present study introduces a novel method that integrates collective mental signals from social networks to predict the collective trajectory of self-harm. The ubiquitous nature of social media renders it a desirable resource for real-time monitoring and forecasting applications, owing to its instantaneous availability and accessibility from any location and at any time. As far as the existing literature is concerned, the utilization of social media data for predicting self-harm rates at the national level has not been investigated.

TABLE 1. Summary of important acronyms used in this paper.

Acronym	Original Term
Mental signals	
ME-Ang	Anger emotion
ME-Dis	Disgust emotion
ME-Fea	Fear emotion
ME-Joy	Joy emotion
ME-Sur	Surprise emotion
ME-Neu	Neutral emotion
MS-Pos	Positive sentiment
MS-Neg	Negative sentiment
MS-Amb	Ambiguous sentiment
MS-Neu	Neutral sentiment
M-ST	Suicidal tendency
M-NST	Non-suicidal tendency
Ground-truth data	
GH-Death	Number of self-harm cases that result in death
GH-Injure	Number of self-harm cases that result in injuries

Mathematically, let X and Y be the sets of mental signal and target variables, respectively. In this research, X comprises normalized mental signals extracted from social media data. These signals are divided into three categories: emotions (ME), sentiments (MS), and suicidal tendencies (M). The emotion signals comprised anger, disgust, fear, joy, surprise, and neutrality. The sentiment signals include positive, negative, ambiguous, and neutral sentiments. The suicidal tendency can be either having a suicidal tendency or not. The selection of these mental signals follows our prior study on the correlation analyses between mental signals extracted from social networks and socioeconomic variables [62]. Table 1 lists the acronyms of all the mental signals used in this research. Furthermore, $Y = \{y_{death}, y_{injury}\}$ comprises the ground-truth actual numbers of deaths (y_{death}) and injuries (y_{injury}), respectively, from self-harm incidents.

Then $X^{(t)} = \langle x_0^{(t)}, x_1^{(t)}, \ldots, x_m^{(t)} \rangle$ and $Y^{(t)} = \langle y_0^{(t)}, y_1^{(t)}, \ldots, y_n^{(t)} \rangle$ represent the vector of m mental signals extracted from social media data and n target attributes, respectively, at time t. Policy-related socioeconomic variables are generally measured on a monthly basis in terms of temporal granularity. However, it is possible for future applications to adjust the frequency granularity to better suit their specific needs. For example, for time-sensitive applications such as predicting daily stock prices, the time steps could be daily, while less time-critical applications such as GDP or agriculture yields can be forecasted in quarters or annuals.

Given a time step t, lag l, and horizon h, our task is to build a forecasting model:

$$\hat{y}^{(t+h)} = f(Y^{[t-l:t-1]}, X^{[t-l:t]}, h) \qquad (1)$$

where $\hat{y}^{(t+h)}$ is the prediction of the target variable $y \in Y$ at time $t+h$ (i.e., h time steps in the future, relative to t), $X^{[t-l:t]}$ denotes the mental signals extracted from social media data collected during time steps $[t-l, t]$, and $Y^{[t-l:t-1]}$ represents the historical values from the time step $t-l$ to $t-1$. Note that $Y^{(t)}$ is not included as an input variable of the model since, in reality, such values may not be available at the current time step. For example, at the end of April 2023, while it makes sense to assume that all the social media data generated in April 2023 can be available for computation, the actual number of self-harm cases reported within this month may need a longer time to aggregate, process, and make available. As a result, the ability to even nowcast ($h = 0$) or hindcast ($h < 0$) has still been deemed valuable in many predictive applications of certain socioeconomic variables whose availability of ground-truth statistics are severely delayed [39], [76].

To encapsulate the entire process of building the forecasting models for population-level self-harm trends, this research presents the *FAST* framework, whose high-level methodology is depicted in Figure 1. First, social media data is automatically collected. Then, each social media message is processed to extract mental signals. These message-wise mental signals are aggregated into temporal periods and normalized with all the messages collected within each period. In parallel, the ground-truth statistics are collected and processed. The extracted, normalized mental signals and ground-truth statistics are temporally aligned and produce a multivariate time series. The time-delay embedding algorithm is performed with different lags and horizons to produce a sample-wise temporal-aware dataset that is compliant for training machine learning regression models. Finally, different regression models are validated for their forecastability at different horizons. The following subsections delve into each of the modules in detail.

A. DATA COLLECTION AND PROCESSING

Social media data is collected from reputable social networking services using their official APIs. Since different social media platforms provide diverse sets of functionalities and features to serve the heterogeneous needs of different user groups and purposes, to minimize any possible assumptions about the available information accompanying social media messages, and allow the proposed framework to generalize across different social media platforms, the smallest unit of social media content is referred to as a message and only comprises timestamped textual content. While collecting social media messages, user-identifiable data, reactions, images, shares, and other platform-specific information are disregarded. The collected messages are stored on local storage for further processing. In this research, a case study of forecasting self-harm trends in Thailand is used to validate the framework. Therefore, social media data used for the experiments comprise Thai tweets collected from the Twitter API.[1]

[1] https://developer.twitter.com/en/docs/twitter-api

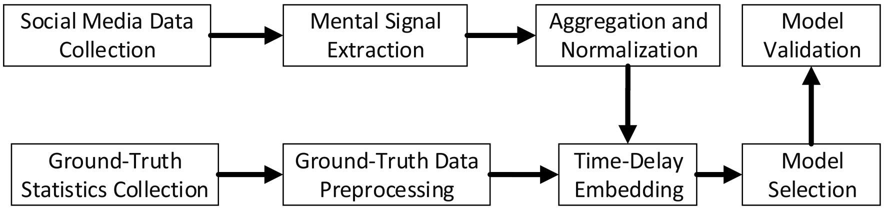

Social Media Data Collection
Mental Signal Extraction
Aggregation and Normalization
Model Validation
Ground-Truth Statistics Collection
Ground-Truth Data Preprocessing
Time-Delay Embedding
Model Selection

FIGURE 1. High-level methodology of the proposed *FAST* framework.

Public ground-truth self-harm statistics can be collected from official administrative organizations. However, different organizations may provide the data in different formats. In our study, the monthly numbers of deaths and injuries from self-harm are collected from the Department of Mental Health of Thailand's Ministry of Public Health.[2] Note that while only the numbers of deaths and injuries are used as the target variables in this study, extending work could easily change the target variables to other socioeconomic variables without impacting the rest of the pipeline.

B. MENTAL SIGNAL EXTRACTION, AGGREGATION, AND NORMALIZATION

Since social media messages comprise only textual information, to capture the population's mental landscape, each message must be quantified for the possible exhibition of mental signals. While one message that shows a mental signal would not be interpretable at the aggregate level, a collective amount of such a signal could be revealing. The problem of mental signal extraction is framed as a multi-label classification task where a given message can be classified into more than one mental signal class. Different classes of mental signals are listed in Table 1. The subsections explain the signal extraction process, followed by aggregation and normalization of these extracted mental signals.

1) MENTAL SIGNAL EXTRACTION

The proposed mental signals can be divided into three categories: emotions, sentiments, and suicidal tendencies. With the availability of sufficient training data, this problem could be easily solved with traditional text classification methods. However, the application of such a supervised method on low-resource language has faced tremendous challenges, one of which is the lack of annotated data. Countering this problem, language-agnostic approaches, such as LaBSE [23], have been proposed that allow the models to train on a source language and apply to a different target language. Specifically, extracting the aforementioned mental signals from social media messages composed in low-resource languages has been investigated in a previous study [62], where LaBSE has been shown effective, surpassing the state-of-the-art machine translation methods. This research, therefore, adopts LaBSE for the mental signal extraction task. The use of language-agnostic models in the proposed framework would promote generalizability to other languages whose corresponding annotated data is insufficient.

TABLE 2. Annotated datasets (composed in English) for training emotion, sentiment, and suicidal tendency signal extraction models.

Mental Signal	Messages	Mental Signal	Messages
Emotion (GoEmotions)		Sentiment (GoEmotions)	
ME-Ang	7,022	MS-Pos	21,733
ME-Dis	816	MS-Neg	11,319
ME-Fea	883	MS-Amb	5,190
ME-Joy	21,119	MS-Neu	16,021
ME-Sad	3,212		
ME-Sur	5,190		
ME-Neu	16,021		
Total	54,263	Total	54,263
Suicidal Tendency (Reddit r/SuicideWatch)			
M-ST	116,037		
M-NST	33,052		
Total	149,089		

Public annotated datasets composed in high-resourced languages (e.g., English) are available for each of the mental signal categories. Specifically, the emotion and sentiment datasets are retrieved from GoEmotions [20], while the annotated suicidal tendency data is collected from r/SuicideWatch subreddit inspired by Shing et al. [86] and later CLPsych 2019 Shared Task [99]. Table 2 summarizes the statistics of these annotated mental signal datasets.

The pre-trained language model utilized in the cross-lingual representation technique is equipped with a document encoder that has the ability to generate language-agnostic representation. When two documents consisting of distinct languages exhibit similar semantics, a language-agnostic language model produces a similar representation. The aforementioned methodology involves semantically embedding texts that are heterogeneous in languages into a shared vector space, thereby facilitating direct comparison between them. LaBSE [23], a language-agnostic deep learning representation model, has been widely utilized in various low-resource language analysis tasks due to its extensive coverage of languages and recent development. For instance, it has been employed in hate speech detection in Marathi [30] and Spanish [79], hope statement detection in Dravidian [89], text reuse discovery in Urdu [61], and sentiment analysis

[2] https://506s.dmh.go.th/Home

in Uyghur [68]. LaBSE was derived from BERT and subsequently fine-tuned to evaluate translation pairs based on their embedding similarity score. Correct sentence pairs were assigned a high score, while incorrect sentence pairs were given a low score. The pre-training process previously required extensive parallel corpora of multiple languages, including texts in both the source and target languages. However, pre-trained models are now readily available to the public for utilization and further investigation.[3] Subsequently, a pre-existing model is fine-tuned in conjunction with a classifier to construct models for the classification of mental signals. These models are developed utilizing the three mental signal datasets previously discussed. The process of fine-tuning a pre-existing model enables it to acquire the necessary representation and discriminative features for specific tasks, ultimately leading to enhanced performance [88]. Finally, the fine-tuned model is used to predict a mental signal for an input text in a local language.

Since the mental signal extraction task is framed as a multi-label classification problem where each mental signal (class) can be treated as a binary classification task – determining whether a message has the target mental signal or not, the evaluation can be carried out using the standard information retrieval protocol for cross-lingual document classification. Specifically, the classification models are trained with source-language documents and evaluated on annotated documents in the target language, using standard precision, recall, and F1 as evaluation metrics [83].

2) MENTAL SIGNAL AGGREGATION AND NORMALIZATION

To comply with the multi-variate time series forecasting methodology, each of these extracted mental signals from social media messages must be aggregated to form a time series. Therefore, the quantified sum of messages that have been normalized and collected during a designated time frame and identified as possessing a mental signal is utilized as a representation of said signal for the specified duration. Mathematically, let D be the set of social media messages, where $D^{(t)} \in D$ represents the set of messages collected during the time period t. Furthermore, for each mental signal $x \in X$, let $D_x^{(t)}$ denote the set of messages exhibiting the mental signal x in the time step t. Then, $x^{(t)}$ is computed as follows:

$$x^{(t)} = \frac{|D_x^{(t)}|}{|D^{(t)}|} \tag{2}$$

The main reason for the normalization is that the sampling ratio of the collected social media messages in each time period can differ. Furthermore, external factors, such as distressful events, feature changes in social networking platforms, and the emergence/shutdown of social networking services, may induce social media users to express their opinions on a specific platform more or less than usual. Therefore, the raw frequencies of mental signals extracted from the collected social media messages could be susceptible to fluctuating volumes of surge or fad discussions. Normalizing the total amount of extracted signals with all the messages collected in each period could help to mitigate such biases.

C. TIME-DELAY EMBEDDING

Once the mental signals are aggregated and aligned with the ground-truth self-harm statistics into a multivariate time series, it is possible to apply the multivariate forecasting framework to evaluate prediction models for future self-harm trends. In this work, in addition to conventional statistical models for multivariate time series such as Vector Autoregression (VAR) and Autoregressive Integrated Moving Average (ARIMA), we also propose to use machine-learning-based multivariate time series forecasting models. The inclusion of machine-learning-based methods is because recent studies have reported that these traditional statistical models have limitations in handling multivariate time series with high dimensionality [48], noises [60], and non-linear patterns [27], while machine learning regression models have been shown to handle large dimensions, resilient against noises and missing values, and capture non-linear relationship [92].

However, traditional machine learning regression models treat data points independently, neglecting their temporal relation, while self-harm trends have been shown to exhibit temporal relationship and seasonality [97]. Therefore, to incorporate temporal dependencies into the feature space, the *time-delay embedding* algorithm [92] is performed on the multivariate time series to produce lag-embedded data points for training machine learning regressors. Mathematically, recall that $X^{(t)} = [x_1^{(t)}\ x_2^{(t)}\ x_3^{(t)}\ \ldots\ x_m^{(t)}]$ and $Y^{(t)} = [y_1^{(t)}\ y_2^{(t)}\ y_3^{(t)}\ \ldots\ y_n^{(t)}]$ represent the snapshots of the mental signals and ground-truth statistics at time t, respectively. Then, let $p^{(t)}(l)$ represent the time-delay embedded version of such an instance with lag l, expressed as follows:

$$p^{(t)}(l) = [\begin{array}{ccccc} x_1^{(t)} & x_1^{(t-1)} & x_1^{(t-2)} & \ldots & x_1^{(t-l)} \\ x_2^{(t)} & x_2^{(t-1)} & x_2^{(t-2)} & \ldots & x_2^{(t-l)} \\ x_3^{(t)} & x_3^{(t-1)} & x_3^{(t-2)} & \ldots & x_3^{(t-l)} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ x_m^{(t)} & x_m^{(t-1)} & x_m^{(t-2)} & \ldots & x_m^{(t-l)} \\ & y_1^{(t-1)} & y_1^{(t-2)} & \ldots & y_1^{(t-l)} \\ & y_2^{(t-1)} & y_2^{(t-2)} & \ldots & y_2^{(t-l)} \\ & \ldots & \ldots & \ldots & \ldots \\ & y_n^{(t-1)} & y_n^{(t-2)} & \ldots & y_n^{(t-l)} \end{array}] \tag{3}$$

The above representation can be shorthanded as:

$$p^{(t)}(l) = X^{[t-l:t]} \oplus Y^{[t-l:t-1]} \tag{4}$$

where $\oplus$ represents the concatenation operator.

The above time-delay representation allows each snapshot of the time series at a given time step to incorporate the previous l values of the mental signal and target variables.

[3] https://tfhub.dev/google/LaBSE/1

For example, with monthly represented time series data, if the current month is $t =$ *April-2023*, then $p^{(t)}(l = 3)$ produces a vector representing the current month's mental signals and the previous values of mental signals and target historical statistics from *January-2023* to *March-2023*. Such a representation also allows the machine learning models to capture the temporal relationship between training samples. It is important to note that $Y^{(t)}$, the values of the target historical statistics at the current time step, are not embedded in the $p^{(t)}(l)$ because socioeconomic statistics are often delayed and, practically, are not available immediately as input features.

D. FORECASTING MODEL SECTION

The time-delay embedding algorithm represents a snapshot of multivariate time-series data with a vector representation incorporating lags. Such a representation also allows linear and non-linear machine learning regression models to train on the same data using the traditional regression methodology. In this research, five machine learning regressors are considered, including Bayesian Ridge [54], Support Vector Machine Regressor with the linear kernel (LinearSVR) [13], XGBoost [18], RandomForest [85], and CatBoost [72]. This research uses the sklearn's implementation[4] of Bayesian Ridge, LinearSVR, and RandomForest, and official releases of XGBoost[5] and CatBoost.[6] The ARIMA model (using the pmdarima[7] implementation) is used as the conventional baseline for comparison and to validate the efficacy of the proposed machine learning-based forecasting models.

E. FORECASTING MODEL EVALUATION

The standard leave-one-out sliding evaluation protocol is used to validate the performance of forecasting models [92]. For each target attribute, y, lag l, and a given time period t, the forecasting model learns the history of mental signal data from the time period $(t - l)$ to (t) and the historical values of y from the time period $(t - l)$ to $(t - 1)$, then predict the target value in the next h time periods. The train-predict cycle continues in this fashion through the rest of the test instances. Note that $h = 0$ is referred to as nowcasting or predicting the value in the current time period. While nowcasting may not be practically useful in time-sensitive applications such as stock price or weather prediction, the ability to predict the current socioeconomic variables is still useful since the availability of these variables is often delayed [26].

Standard evaluation metrics for forecasting models are used, including Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE). The Mann-Whitney U test is used to quantify the statistical difference between the baseline (ARIMA) and other machine learning regressors. Furthermore, the *error slope* represents the slope of the linear line that fits the absolute percentage errors as the function of time. A negative error slope can be used as an indicator that the forecasting error generally becomes smaller as the model learns more data.

IV. EXPERIMENT, RESULTS, AND DISCUSSION

It is important that the proposed framework is evaluated on real-world data. This section first discusses social media data and self-harm statistics used in this research. Then, the evaluation performance of the mental signal extraction task, replicated from [62], is briefly explained. The aggregate mental signals and ground-truth statistics were then used to validate forecasting models. The best forecaster and the baseline ARIMA models were then used to validate forecastability by varying different horizons. Finally, the impact of various feature types on forecasting performance is discussed.

A. DATASETS

This research aims to determine if mental signals extracted from social media data could be used to forecast the number of death and injury cases from self-harm at the national level. A case study of social media and ground-truth statistics of self-harm cases in Thailand was used in this study. Note that since the method for extracting mental signals is language-agnostic, and the proposed forecasting approach is applicable as long as the data is in the form of multivariate time series, the proposed framework could easily be generalized and adopted in different linguistic and geographical contexts.

Two datasets are used in this research. The first comprises tweets in Thailand randomly collected using the Twitter API from October 2017 to January 2021, totaling roughly 4.9 million tweets. For generalizability across other social media platforms, only the timestamp and textual information were retained from each tweet. The second dataset comprises the ground-truth historical statistics of monthly cases of death and injury from self-harm made publicly available by the Department of Mental Health, Ministry of Public Health of Thailand. Figure 2 illustrates the monthly numbers of tweets (bar chart with the right Y-axis) and the numbers of reported deaths and injuries from self-harm (line chart with the left Y-axis). It is interesting to note the rise in both death and injury cases from September 2019 to October 2019, which could be due to the changes in reporting system as a result of the fiscal year transition (a fiscal year in Thailand begins from October to September).

B. SOCIAL MENTAL SIGNAL EXTRACTION

Mental signals are extracted from each tweet. Extracting each mental signal was treated as a binary classification task, where a message was classified whether having such a signal or not. In this work, following Noraset et al. [62]'s, three categories of mental signals were considered: sentiment, emotion, and suicidal tendency. The list of individual signals belonging to each category is provided in Table 1.

In this work, the language-agnostic mental signal extraction approach developed as part of our previous work [62]

[4] https://scikit-learn.org/
[5] https://xgboost.readthedocs.io
[6] https://catboost.ai/
[7] https://pypi.org/project/pmdarima/

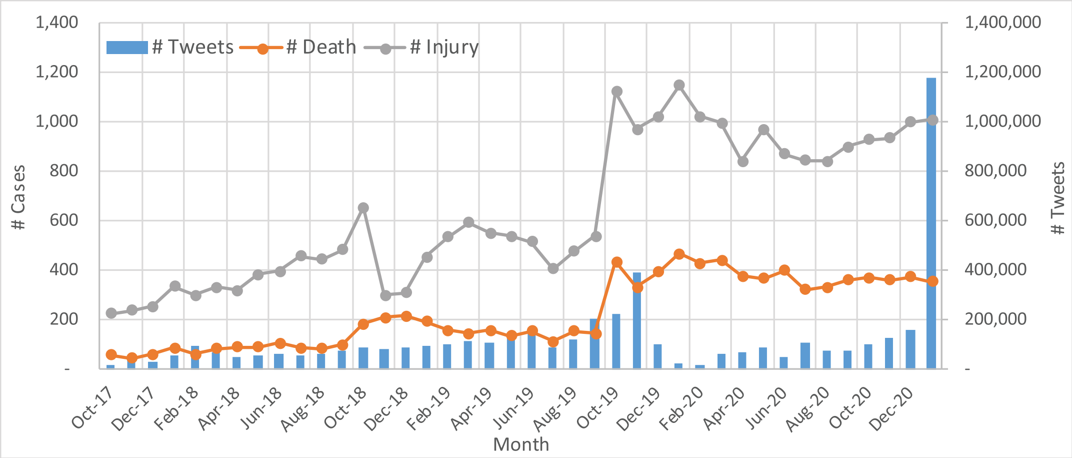


FIGURE 2. Statistics of the collected datasets, including the monthly numbers of tweets (second axis) and cases of self-harm that result in death and injury.

TABLE 3. Classification performance of the mental signal extraction using LaBSE on Thai tweets, as reported in [62].

Type	Class	Precision	Recall	F1
Sentiment	MS-Pos	0.57	0.77	0.66
	MS-Neg	0.7	0.83	0.76
	MS-Amb	0.87	0.46	0.6
	MS-Neu	0.33	0.92	0.49
	Macro Avg	0.62	0.74	0.63
Emotion	ME-Ang	0.52	0.48	0.5
	ME-Dis	0.98	0.28	0.44
	ME-Fea	0.93	0.88	0.9
	ME-Joy	0.55	0.96	0.7
	ME-Sad	0.78	0.89	0.83
	ME-Sur	0.6	0.88	0.71
	ME-Neu	0.59	0.79	0.68
	Macro Avg	0.71	0.74	0.68
Suicidal Tendency	M-ST	0.75	0.89	0.81
	M-NST	0.87	0.7	0.78
	Macro Avg	0.81	0.8	0.8

was adopted since the Thai tweets were also used as their case study. Their mental extraction approach comprises three multi-label classifiers trained using LaBSE on the source English language datasets, previously discussed in Section III-B1. The evaluation was conducted by testing the trained models on the target manually annotated Thai tweets, where the classification performance is quoted in Table 3. Interested readers should refer to the original work for more detail.

The mental signal extractors were applied to the collected tweets. The tweets that display each signal are aggregated monthly and normalized by the total number of tweets in that month. The monthly normalized mental signals collected from our social media data are shown in Figure 3. It is worth noting that the suicidal tendency (M-ST) increased dramatically in January 2020. That was also when the first COVID-19 case was identified in Thailand [43], which might have caused widespread dread and worry [8].

C. FORECASTING MODEL SELECTION

Once the mental signals and ground-truth self-harm statistics were formulated as multivariate time series, the time-delay embedding algorithm (Section III-C) was performed to generate an instance-like, temporal-embedded dataset for training machine learning regressors. A selection of regression models listed in Section III-D was tested along with the baseline ARIMA model using the leave-one-out sliding evaluation protocol discussed in Section III-E. The last 12 months of data (February 2020 - January 2021) were allocated as the test data.

The objective of the model selection experiment was to identify the best regression model to conduct further analyses. In this experiment, the horizon is fixed at $h = 0$, while the lags were varied from 0 - 12 (i.e., $l \in \{0, 1, 2, \ldots, 12\}$). The reason for varying the lag is that different regression models may have different capacities to capture both the amount and seasonality of longitudinal data. Table 4 summarizes the forecasting performance ($h = 0$) of different forecasting models with their best-performing lags on both the death and injury tasks against that of the ARIMA baseline. The Mann-Whitney U test was conducted between each model's predictions against the baseline ARIMA, where the p-values are also reported. Finally, the error slope represents the overall trend of the error when the model is trained with more data. A negative error slope indicates that the prediction errors become smaller as a function of time, implying that the model could perform better if more training data were available.

Fixing $h = 0$, XGboost with $l = 1$ performed best in all evaluation metrics, outperforming ARIMA by 65.98% (MAPE reduces from 19.37% to 6.59%) and 51.18% (MAPE reduces from 12.70% to 6.20%) for the death and injury forecasting tasks, respectively. Using $\alpha = 0.05$, the statistic test shows that XGBoost's predictions are statistically different from those of the baseline in both tasks. It is worth noting that most of the machine learning based regression models outperformed ARIMA in the death forecasting task, while

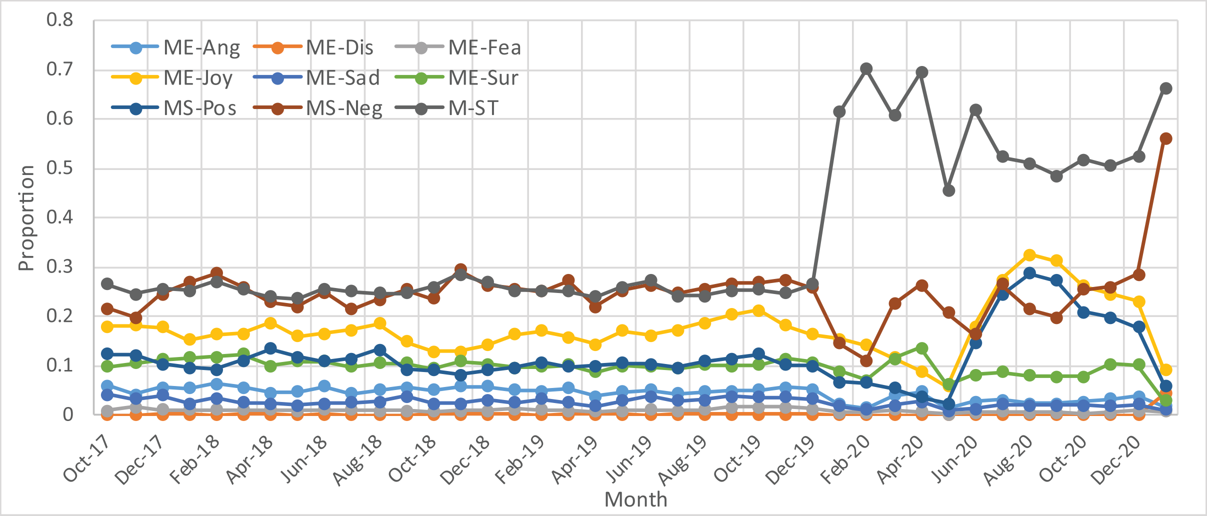


FIGURE 3. Visualization of the normalized aggregate mental signal computed from tweets collected each month.

TABLE 4. Comparison of the prediction performance of different regressors and their best-performing lags, keeping horizon fixed at 0 (i.e., $h = 0$). Bold-italic figures represent the best performance with respect to each evaluation metric.

Variable	Regressor	Best Lag	MAE	RMSE	MAPE	p-value	Error Slope
Death	ARIMA	12	69.316	85.951	19.37%	-	0.021
	Bayesian Ridge	3	41.223	49.807	10.78%	0.118	-0.006
	Linear SVR	3	65.256	85.561	17.98%	0.419	0.010
	XGBoost	1	***23.989***	***29.911***	***6.59%***	0.034	-0.004
	Random Forest	0	45.180	71.299	11.14%	0.056	-0.006
	CatBoost	1	61.045	83.370	15.51%	0.375	-0.008
Injury	ARIMA	12	118.067	145.864	12.70%	-	0.006
	Bayesian Ridge	3	104.023	127.270	11.33%	0.420	-0.011
	Linear SVR	12	128.397	169.058	13.89%	0.420	-0.004
	XGBoost	1	***57.670***	***66.369***	***6.20%***	0.034	0.002
	Random Forest	12	82.018	115.257	8.56%	0.130	-0.015
	CatBoost	1	131.056	167.577	13.63%	0.354	-0.010

only CatBoost was inferior to ARIMA in the injury forecasting task. This could be due to the fact that the input features are all numeric (i.e., not categorical), making CatBoost not a suitable choice for this problem. Among the machine learning regressors, besides CatBoost, Linear SVR also performs worst than others in both tasks. This could shed light on the nature of the relationship between temporal mental signals and the target variables, which may not be linear. Another interesting point is that different models require different lags to perform optimally. For example, the best forecaster, XGBoost, only requires a lag of one (i.e., this and previous months of mental signals, and only the previous month of ground-truth statistics) to make an optimal prediction of the current target's value. However, the baseline ARIMA model needs the largest experimented lag of 12 to perform optimally.

Figure 4 plots the predictions at $h = 0$ of the XGBoost (dashed green line) and ARIMA (dashed green line) models against the actual values of the death (left) and injury (right) cases. Visually, the predictions from XGBoost have a smaller absolute error each month. Figure 5 plots the absolute percentage errors (APE) of the ARIMA (orange line) and XGBoost (green line) as the function of time for the death (left) and injury (right) tasks, respectively. In the death forecasting task, the error trend of ARIMA is increasing while that of the XGBoost fluctuates but steadily declines as the model learns from more data. Such an error trend indicates that the XGBoost model for predicting death cases could still improve with more data. However, in the injury prediction task, the overall trend of APE of both the ARIMA and XGBoost models increases. However, the XGBoost's error trend for injury prediction does not seem to fluctuate much compared to the death prediction task, suggesting that the model could already have stabilized.

D. IMPACT OF HORIZONS ON FORECASTABILITY

The previous experiment held the horizon constant at zero while varying the forecasting models. The best model was chosen for the forecastability evaluation presented in this section, compared to the baseline ARIMA model. From the previous section, XGBoost was determined as the best

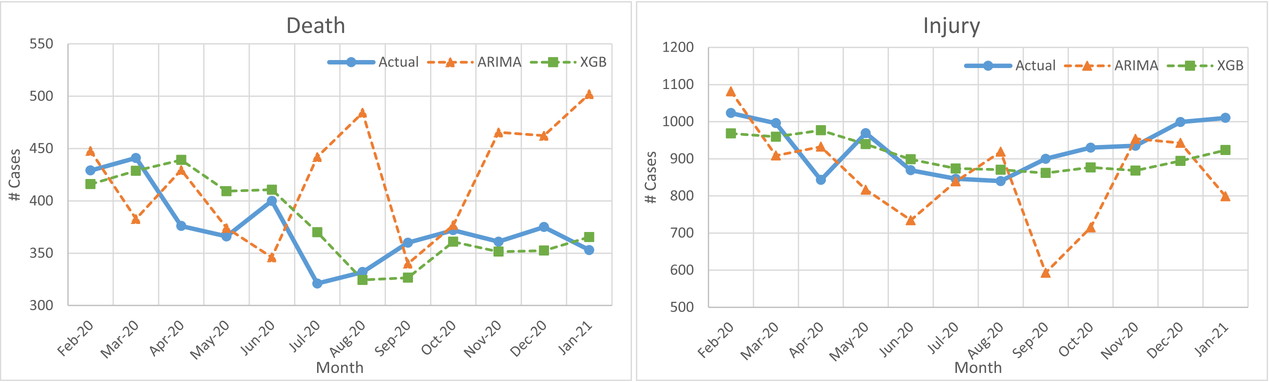


FIGURE 4. Visualization of prediction from ARIMA and XGBoost regressors with $h = 0$, against the actual values, on both the Death (left) and Injury (right) variables.

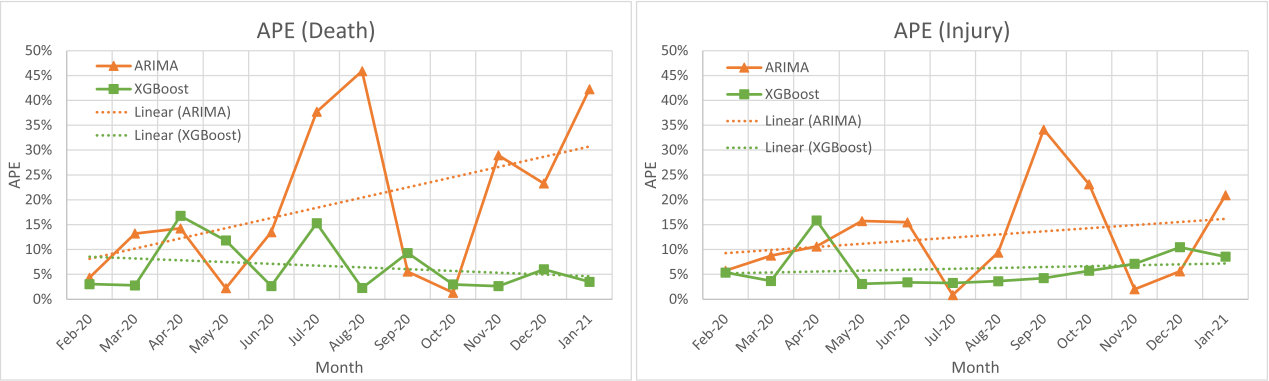


FIGURE 5. Visualization of monthly average percentage errors from ARIMA and XGBoost regressors with $h = 0$ on both the Death (left) and Injury (right) variables.

forecaster for both the death and injury forecasting tasks due to its lowest average MAPE compared to others. In this section, horizons are varied from zero to six (i.e., $h \in \{0, 1, 2, 3, 4, 5, 6\}$) to assess the model's performance when predicting different future time steps. For example, the model with $h = 3$ predicts the target value three months later. For each horizon, the lags are varied again (i.e., $l \in \{0, 1, 2, 3, \ldots, 12\}$) to find the optimal one for each horizon.

Table 5 summarizes the forecasting performance of the ARIMA (baseline) and XGBoost models at different horizons for the death and injury tasks. The Mann-Whitney U test was performed to compare the predictions of the ARIMA and XGBoost models at each horizon, whose p-values are also reported. In both the death and injury tasks, the forecasting errors in terms of MAPE generally rise as h increases for both the ARIMA and XGBoost models. This phenomenon is expected since predicting too far away into the future result in lower prediction confidence. At each horizon, XGBoost has better forecasting performance in terms of MAPE compared to ARIMA. On average, XGBoost yields the MAPE of 13.27% and 10.15%, outperforming ARIMA by 43.56% and 36.48% on the death and injury forecasting tasks, respectively. Furthermore, it is observed that the forecasting errors (MAPE) drop at $h = 0, 1,$ and 3 for XGBoost in both the death and injury tasks. This could mean that the forecasting could exhibit quarterly (3-month) seasonality. However, further statistical proof is needed to conclude such a phenomenon.

E. IMPACT OF DIFFERENT FEATURES ON FORECASTABILITY

Since the input features of the forecasting models comprise different types of mental signals and historical statistics, a natural question could arise as to how each feature group contributes to the model's forecastability. In this section, the XGBoost model was chosen to investigate this aspect. The features were prepared in six groups, including 1) historical statistics of the target variables only, 2) the emotion signals only, 3) the sentiment signals only, 4) suicidal tendency only, 5) all the mental signals (i.e., mental + sentiment + suicidal tendency), and 6) all the features combined. The insights from this section's analysis could guide the implementation of the self-harm forecasting system with uncertain availability of data. For example, historical statistics of self-harm cases may not be readily accessible for several months due to a variety of factors, such as modifications to data governance policies or delays in data processing and aggregation. In this scenario, only social media data would be available for the system to update the prediction models. Therefore, it is essential to be

TABLE 5. Forecasting performance of ARIMA and XGBoost at different horizons (i.e., $h = \{0, 1, 2, 3, 4, 5, 6\}$).

Variable	Horizon	ARIMA (Baseline)					XGBoost					
		Best Lag	MAE	RMSE	MAPE	Error Slope	Best Lag	MAE	RMSE	MAPE	p-value	Error Slope
Death	0	12	69.316	85.951	19.37%	0.184	1	23.989	29.911	6.59%	0.034	-0.275
	1	12	62.784	83.627	17.80%	0.225	1	54.764	95.142	14.21%	0.332	-0.271
	2	12	107.865	124.803	28.52%	-0.190	1	70.901	120.471	17.53%	0.030	0.130
	3	9	98.520	116.626	26.42%	-0.266	9	38.171	59.968	10.89%	0.004	-0.217
	4	9	107.593	130.987	30.54%	0.300	9	65.219	113.423	17.00%	0.023	0.051
	5	9	93.230	102.663	24.80%	-0.271	0	64.369	88.671	17.33%	0.056	0.160
	6	12	64.178	74.231	17.11%	-0.160	12	34.499	45.443	9.34%	0.020	0.030
	Average	-	86.212	102.698	23.51%	-0.026	-	50.273	79.004	13.27%	0.071	-0.056
Injury	0	12	118.067	145.864	12.70%	-0.401	0	57.670	66.369	6.20%	0.034	0.002
	1	12	169.038	207.951	18.23%	-0.732	1	113.216	184.901	11.56%	0.087	-0.025
	2	12	161.757	204.599	17.43%	-0.352	12	114.227	175.627	11.97%	0.170	-0.012
	3	9	168.086	236.038	18.03%	-1.070	9	85.759	112.783	9.42%	0.143	-0.002
	4	12	164.322	195.911	17.66%	-0.933	12	122.510	174.879	13.04%	0.143	-0.012
	5	12	136.376	163.939	14.52%	-0.876	12	103.940	143.429	11.42%	0.185	0.000
	6	12	124.776	153.319	13.31%	-0.545	12	68.817	77.387	7.47%	0.118	-0.004
	Average	-	148.917	186.803	15.98%	-0.701	-	95.163	133.625	10.15%	0.126	-0.008

able to anticipate the performance of models when certain features become unavailable.

The XGBoost model was evaluated by training with each different feature type to predict the target attributes at horizons $h = 0$ and 3. These horizons were used as representatives to investigate the impact of different feature types on the models' ability to nowcast ($h = 0$) and forecast ($h = 3$), respectively. Table 6 reports the results for both the death and injury tasks. Δ MAPE denotes the relative difference of MAPE with respect to using historical statistics alone. Note that the MAPE is 100% for the injury nowcasting prediction task using only historical statistics because the best lag for XGBoost ($h = 0$) from the previous experiment was 0, meaning that the model did not have any historical statistics nor mental signals as input, resulting in predicting the default values, i.e., zeroes.

The first point to note is that models trained with only mental signals (i.e., Social-Combined) have better predictions than those trained only with the historical statistics for $h = 0$ death, $h = 3$ death, and $h = 0$ injury prediction, over-performing the model trained only with historical statistics by 23%, 21.74%, and 88.17% in terms of MAPE, respectively. However, for $h = 3$ injury prediction, the model trained with mental signals only has poorer performance than using only the historical statistics. Second, among the mental signals (i.e., emotion, sentiment, and suicidal tendency), models trained with only the sentiment features perform best for $h = 0$ death, $h = 3$ death, and $h = 3$ injury prediction tasks, while the suicidal tendency features outperformed other mental signals in the $h = 0$ injury prediction. Third, combining all the historical statistics and mental signal features yields the best performance for nowcasting death cases, resulting in a MAPE of 6.59%. However, incorporating historical statistics into the feature space for forecasting the death cases three months ahead ($h = 3$) does not affect the performance much (MAPE = 10.89% using all features vs. MAPE = 10.88% using only mental signal features). It is interesting to note that adding mental signals worsens the three-month forecasting of injury cases, where the historical statistics alone already yields a MAPE of 7.11%, but using all the features increases the MAPE to 9.42% (32.45% performance drop). To summarize, the integration of mental signals derived from social media is demonstrated to enhance the forecasting accuracy of death cases from self-harm in our particular case study. The utilization of mental signals in the injury forecasting task has been observed to enhance the nowcasting ($h = 0$) performance. However, additional analyses are required to determine the extent of their contribution toward the ability to forecast the trends of self-harm injuries.

Table 7 lists the top ten features from the XGBoost forecasting models at horizons $h = 0$, 1, 2, and 3 for both the death and injury tasks. Features starting with GH, ME, MS, and M represent ground-truth statistics, emotion, sentiment, and suicidal tendency groups, respectively. The parenthesis behind each feature denotes its lag. For example, *MS-Pos(t-1)* represents the positive sentiment extracted from social media in the previous month ($t - 1$). In the context of both the death and injury forecasting tasks, it has been observed that the predominant top-1 features across various horizons are the ground-truth historical statistics, denoted by *GH-*. It is not unexpected that the default values of the target attributes could be determined by their prior values, similar to the bias terms in linear regression. Additional features may also contribute to the models, thereby enhancing their predictive accuracy. Furthermore, for the prediction of the death cases, using $h = 0$ as an example, most of the top features are sentiment signals (i.e., *MS-*), while previous ground-truth statistics (i.e., *GH-*) are among the top features for the injury prediction task.

F. LIMITATIONS

While the experiment results showed that mental signals extracted from large-scale social networks could improve the forecasting of self-harm trends using a case study of Thailand, the proposed framework still has room for improvement.

TABLE 6. Forecasting performance of XGBoost regressor on the Death and Injury variables, trained with different types of features.

Variable	Feature Type	h=0			h=3		
		MAPE	Δ MAPE	p-value	MAPE	Δ MAPE	p-value
Death	Historical Statistics*	9.23%	-	-	13.90%	-	-
	Emotion	10.78%	-16.80%	0.465	9.85%	29.15%	0.17
	Sentiment	8.43%	8.65%	0.465	10.76%	22.60%	0.218
	Suicidal Tendency	8.57%	7.10%	0.488	13.09%	5.83%	0.354
	Social-Combined	7.11%	23.00%	0.201	10.88%	21.74%	0.17
	All	6.59%	28.61%	0.292	10.89%	21.63%	0.185
Injury	Historical Statistics*	100.00%	-	-	7.11%	-	-
	Emotion	11.58%	88.42%	< 0.001	21.42%	-201.27%	0.007
	Sentiment	10.83%	89.17%	< 0.001	13.55%	-90.58%	0.063
	Suicidal Tendency	7.30%	92.70%	< 0.001	15.07%	-112.00%	0.013
	Social-Combined	11.83%	88.17%	< 0.001	17.25%	-142.71%	0.034
	All	6.20%	93.80%	< 0.001	9.42%	-32.45%	0.201

TABLE 7. Top ten features ranked by XGBoost feature importance on both Death and Injury variables trained to forecast at different horizons.

Top Feature	Death				Injury			
	h = 0	h = 1	h = 2	h = 3	h = 0	h = 1	h = 2	h = 3
1	GH-Injure(t-1)	GH-Injure(t-1)	GH-Injure(t-1)	GH-Death(t-9)	ME-Dis(t)	GH-Injure(t-1)	GH-Injure(t-10)	GH-Injure(t-9)
2	MS-Neu(t-1)	ME-Joy(t-1)	ME-Joy(t)	ME-Ang(t-5)	ME-Neu(t)	ME-Joy(t-1)	GH-Injure(t-9)	MS-Amb(t-6)
3	MS-Amb(t-1)	MS-Amb(t)	MS-Pos(t)	MS-Amb(t-7)	ME-Fea(t)	ME-Neu(t-1)	GH-Death(t-6)	ME-Joy(t-8)
4	MS-Pos(t-1)	MS-Pos(t)	MS-Amb(t)	ME-Sad(t-1)	MS-Amb(t)	GH-Injure(t-1)	ME-Joy(t-9)	GH-Death(t-3)
5	ME-Neu(t)	GH-Death(t-1)	ME-Sur(t)	MS-Amb(t-8)	MS-Neu(t)	MS-Amb(t)	GH-Death(t-5)	ME-Joy(t-2)
6	ME-Sur(t-1)	ME-Neu(t)	MS-Amb(t-1)	MS-Amb(t-9)	ME-Ang(t)	ME-Fea(t-1)	GH-Death(t-4)	M-NST(t-1)
7	ME-Neu(t-1)	ME-Joy(t)	ME-Dis(t-1)	ME-Ang(t-9)	M-NST(t)	MS-Amb(t-1)	MS-Amb(t-7)	GH-Death(t-6)
8	M-NST(t)	MS-Amb(t-1)	M-NST(t)	ME-Sur(t-7)	ME-Sur(t)	M-NST(t)	ME-Joy(t-1)	GH-Injure(t-8)
9	MS-Amb(t)	ME-Dis(t-1)	ME-Fea(t)	ME-Neu(t-9)	MS-Pos(t)	ME-Ang(t-1)	MS-Amb(t-11)	ME-Sur(t-6)
10	ME-Joy(t-1)	M-NST(t-1)	MS-Pos(t-1)	MS-Neg(t-1)	MS-Neg(t)	ME-Sur(t)	ME-Sad(t-2)	M-NST(t-3)

First, only conventional machine learning regression algorithms were validated as the forecasters. However, the advent of deep learning technologies has given birth to many sequence models that can be used directly as multi-variate time series forecasters, such as GatedTabTransformer [19] and Time Series Transformer [98]. These deep learning algorithms were not primarily explored in this research due to the limited longitudinal data in our study, which could be insufficient for deep learning models to work well [7]. Our future direction intends to explore these deep-learning-based time series models. Furthermore, the experimental results presented in this paper pertain to the case study of monthly self-harm reported incidents in Thailand, using Thai tweets as the social media data source. However, since minimal assumptions were made about these two sources of data, including the frequencies of data availability and the languages used to compose social media messages, the proposed framework could be generalized across many forms of historical statistics and social media platforms. Finally, the experiments presented in this paper only used Twitter as the only social media data source. However, combining information from various other online media sources, such as Facebook and online news articles, could provide richer information to the forecasting models. We aim to further investigate this direction as well.

V. SOCIETAL IMPLICATIONS

The proposed *FAST* framework relies on publicly available data to extract public mental signals that are used to fuel the machine-learning-based models to nowcast and forecast national-level self-harm trends. If the framework were to be adopted for real-world applications, societal ramifications could certainly arise.

A. COSTS TO THE PEOPLE

Two changes will happen if a government agency adopts the methodology to monitor self-harm trends and advise policymakers. To begin, developing the framework into a system requires implementation, computer infrastructure, and maintenance. Furthermore, gathering large amounts of social media data may result in data access fees. Second, using the implemented system, if policymakers anticipate significant rises in self-harm patterns, then prompt actions will be taken to deal with these expected instances. Some of these initiatives would almost certainly result in the implementation of on-the-spot mental health services or hotlines. All of these changes need financial resources, which will most likely be provided by tax revenues. While it was not our intention for society to shoulder this burden, the adopting government would have to weigh the expenses of establishing the system, as well as the procedural costs that follow, with the benefits that people would receive.

B. PRIVACY

Although the framework solely relies on non-personally identifiable information extracted from publicly available social media data, there is a possibility of privacy concerns emerging. Individuals who utilize social media, particularly in developing nations, may lack sufficient training in digital literacy and may not possess a comprehensive understanding of the potential ramifications of their actions on social networking platforms. Certain communications within an individual's social network may be inadvertently or deliberately disclosed to the public, unbeknownst to them that such information may also be subject to government surveillance. To address this concern, it is suggested that social media platforms adopt measures that encourage their users to exercise caution and verify the accuracy of their content prior to publishing it. In addition, it is recommended that the government collaborate with social media platforms to promote the potential utility of their publicly available data. Failure to take appropriate actions could potentially incite unrest and anxiety among people, possibly leading to conjecture regarding the government's involvement in cyberespionage against its own citizens.

C. ABUSE

The proposed framework has been designed with the aim of serving the social good. However, the framework's potential to extract public sentiment from online social networks and forecast self-harm trends may have both positive and negative implications. Given that the data utilized within this framework is publicly available, it is conceivable that individuals with criminal intent may also exploit it for their own benefit. In the event of an anticipated increase in self-harm trends, for example, individuals with malicious intent may seek to exploit these vulnerable populations by engaging in the sale of illicit substances or offering fraudulent consultation services. As another example, if self-harm trends are anticipated to increase after the implementation of stringent policies (e.g., those intended to combat the COVID-19 pandemic), this could potentially create a sense of public apprehension, which opposing political parties may exploit to criticize the government's management of the issue.

VI. CONCLUSION

Self-harm refers to acts of self-poisoning or self-injury that, regardless of intention, result in death or non-fatal injuries. Research has shown the rising trends of self-harm associated with the advent of technologies and rapid urbanization in developing countries. The ability to forecast or even nowcast national-level self-harm trends could prove crucial for policymakers and public health stakeholders in implementing timely procedures to neutralize the root causes or prevent these anticipated tragedies. While previous work has used historical statistics to forecast population-level self-harm, in certain countries, such ground-truth previous statistics may be unavailable, insufficient, and delayed, hindering the timely monitoring of the self-harm landscape for proactive policy-making purposes. This paper proposed *FAST*, a framework for forecasting population-level self-harm trends using mental signals extracted from large-scale social media data. A case study using a set of 12 mental signals extracted from tweets to improve the forecastability of the death and injury cases from self-harm in Thailand illustrated the applicability of the proposed method that outperforms the traditional ARIMA baseline by 43.56% and 36.48% on average in terms of MAPE, respectively. To the best of our knowledge, we are the first to investigate using aggregate social media data to improve nowcasting and forecasting self-harm cases at the national level. While the experimental results are promising, there is still room for improvement. Future research could investigate the use of other online media, such as news articles, other social media platforms, or types of media, such as videos or photos in addition to texts, as well as deep learning forecasting techniques. Moreover, it would be worthwhile to investigate the potential for predicting incidents of self-harm on a more granular, such as regional or demographic-specific, scale. This could facilitate the development of tailored self-harm management approaches that are appropriate for specific localities and demography.

REFERENCES

[1] (May 2023). *Suicide Rates for Girls Are Rising. Are Smartphones to Blame?* [Online]. Available: https://www.economist.com/graphic-detail/2023/05/03/suicide-rates-for-girls-are-rising-are-smartphones-to-blame

[2] J. Adam-Troian and T. Arciszewski, "Absolutist words from search volume data predict state-level suicide rates in the United States," *Clin. Psychol. Sci.*, vol. 8, no. 4, pp. 788–793, Jul. 2020.

[3] A. E. Aiello, A. Renson, and P. Zivich, "Social media-and Internet-based disease surveillance for public health," *Annu. Rev. Public Health*, vol. 41, p. 101, Apr. 2020.

[4] M. Akyuz and C. Karul, "The effect of economic factors on suicide: An analysis of a developing country," *Int. J. Hum. Rights Healthcare*, Jul. 2022.

[5] S. Z. Alavijeh, F. Zarrinkalam, Z. Noorian, A. Mehrpour, and K. Etminani, "What users' musical preference on Twitter reveals about psychological disorders," *Inf. Process. Manage.*, vol. 60, no. 3, May 2023, Art. no. 103269.

[6] A. Aldayel and W. Magdy, "Stance detection on social media: State of the art and trends," *Inf. Process. Manage.*, vol. 58, no. 4, Jul. 2021, Art. no. 102597.

[7] P. Angelov and A. Sperduti, "Challenges in deep learning," in *Proc. 24th Eur. Symp. Artif. Neural Netw. (ESANN)*, 2016, pp. 489–496.

[8] A. Apisarnthanarak, P. Apisarnthanarak, C. Siripraparat, P. Saengaram, N. Leeprechanon, and D. J. Weber, "Impact of anxiety and fear for COVID-19 toward infection control practices among Thai healthcare workers," *Infection Control Hospital Epidemiol.*, vol. 41, no. 9, pp. 1093–1094, Sep. 2020.

[9] S. Arunpongpaisal, S. Assanagkornchai, V. Chongsuvivatwong, and N. Jampathong, "Time-series analysis of trends in the incidence rates of successful and attempted suicides in Thailand in 2013–2019 and their predictors," *BMC Psychiatry*, vol. 22, no. 1, pp. 1–11, Aug. 2022.

[10] J. M. Barros, R. Melia, K. Francis, J. Bogue, M. O'Sullivan, K. Young, R. A. Bernert, D. Rebholz-Schuhmann, and J. Duggan, "The validity of Google trends search volumes for behavioral forecasting of national suicide rates in Ireland," *Int. J. Environ. Res. Public Health*, vol. 16, no. 17, p. 3201, Sep. 2019.

[11] B. E. Belsher, D. J. Smolenski, L. D. Pruitt, N. E. Bush, E. H. Beech, D. E. Workman, R. L. Morgan, D. P. Evatt, J. Tucker, and N. A. Skopp, "Prediction models for suicide attempts and deaths: A systematic review and simulation," *JAMA Psychiatry*, vol. 76, no. 6, pp. 642–651, 2019.

[12] L. Braghieri, R. Levy, and A. Makarin, "Social media and mental health," *Amer. Econ. Rev.*, vol. 112, no. 11, pp. 3660–3693, 2022.

[13] R. G. Brereton and G. R. Lloyd, "Support vector machines for classification and regression," *Analyst*, vol. 135, no. 2, pp. 230–267, 2010.

[14] J. A. Bridge, M. Olfson, J. M. Caterino, S. W. Cullen, A. Diana, M. Frankel, and S. C. Marcus, "Emergency department management of deliberate self-harm: A national survey," *JAMA Psychiatry*, vol. 76, no. 6, pp. 652–654, 2019.

[15] L. Cao, H. Zhang, and L. Feng, "Building and using personal knowledge graph to improve suicidal ideation detection on social media," *IEEE Trans. Multimedia*, vol. 24, pp. 87–102, 2022.

[16] M. K. Y. Chan, H. Bhatti, N. Meader, S. Stockton, J. Evans, R. C. O'Connor, N. Kapur, and T. Kendall, "Predicting suicide following self-harm: Systematic review of risk factors and risk scales," *Brit. J. Psychiatry*, vol. 209, no. 4, pp. 277–283, Oct. 2016.

[17] Y. S. Chang and J. Lee, "Is forecasting future suicide rate possible?—Application of experience curve," *Eng. Manag. Res.*, vol. 1, no. 1, p. 10, 2012.

[18] T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, and H. Cho, "XGBoost: Extreme gradient boosting," R Package Version 0.4-2, Microsoft, Tech. Rep., 2015, pp. 1–4, vol. 1, no. 4.

[19] R. Cholakov and T. Kolev, "The GatedTabTransformer. An enhanced deep learning architecture for tabular modeling," 2022, *arXiv:2201.00199*.

[20] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi, "GoEmotions: A dataset of fine-grained emotions," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, Jul. 2020, pp. 4040–4054. [Online]. Available: https://aclanthology.org/2020.acl-main.372, doi: 10.18653/v1/2020.acl-main.372.

[21] J. B. Edgcomb, T. Shaddox, G. Hellemann, and J. O. Brooks, "Predicting suicidal behavior and self-harm after general hospitalization of adults with serious mental illness," *J. Psychiatric Res.*, vol. 136, pp. 515–521, Apr. 2021.

[22] L. Favril, R. Yu, K. Hawton, and S. Fazel, "Risk factors for self-harm in prison: A systematic review and meta-analysis," *Lancet Psychiatry*, vol. 7, no. 8, pp. 682–691, Aug. 2020.

[23] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang, "Language-agnostic BERT sentence embedding," 2020, *arXiv:2007.01852*.

[24] H. Fliege, J.-R. Lee, A. Grimm, and B. F. Klapp, "Risk factors and correlates of deliberate self-harm behavior: A systematic review," *J. Psychosomatic Res.*, vol. 66, no. 6, pp. 477–493, Jun. 2009.

[25] A. Franzén, "Big data, big problems: Why scientists should refrain from using Google trends," *Acta Sociologica*, pp. 1–5, Jan. 2023.

[26] A. Froidevaux, J. Macalos, I. Khalfoun, M. Deffrasnes, S. d'Orsetti, N. Salez, and A. Sciberras, "Leveraging alternative data sources for socio-economic nowcasting," in *Proc. Conf. Inf. Technol. Social Good*, Sep. 2022, pp. 345–352.

[27] R. Gao, O. Duru, and K. F. Yuen, "High-dimensional lag structure optimization of fuzzy time series," *Expert Syst. Appl.*, vol. 173, Jul. 2021, Art. no. 114698.

[28] M. George, "The importance of social media content for Teens' risks for self-harm," *J. Adolescent Health*, vol. 65, no. 1, pp. 9–10, Jul. 2019.

[29] N. A. Ghani, S. Hamid, I. A. T. Hashem, and E. Ahmed, "Social media big data analytics: A survey," *Comput. Hum. Behav.*, vol. 101, pp. 417–428, Dec. 2018.

[30] A. Glazkova, M. Kadantsev, and M. Glazkov, "Fine-tuning of pre-trained transformers for hate, offensive, and profane content detection in English and Marathi," 2021, *arXiv:2110.12687*.

[31] C. R. Glenn, E. M. Kleiman, J. Kellerman, O. Pollak, C. B. Cha, E. C. Esposito, A. C. Porter, P. A. Wyman, and A. E. Boatman, "Annual research review: A meta-analytic review of worldwide suicide rates in adolescents," *J. Child Psychol. Psychiatry*, vol. 61, no. 3, pp. 294–308, Mar. 2020.

[32] S. D. Gollapalli, G. A. Zagatti, and S.-K. Ng, "Suicide risk prediction by tracking self-harm aspects in tweets: NUS-IDS at the CLPsych 2021 shared task," in *Proc. 7th Workshop Comput. Linguistics Clin. Psychol., Improving Access*, 2021, pp. 93–98.

[33] K. L. Gratz, "Risk factors for and functions of deliberate self-harm: An empirical and conceptual review," *Clin. Psychol., Sci. Pract.*, vol. 10, no. 2, pp. 192–205, 2003.

[34] K. Hawton, D. Zahl, and R. Weatherall, "Suicide following deliberate self-harm: Long-term follow-up of patients who presented to a general hospital," *Brit. J. Psychiatry*, vol. 182, no. 6, pp. 537–542, Jun. 2003.

[35] K. Hawton, K. E. Saunders, and R. C. O'Connor, "Self-harm and suicide in adolescents," *Lancet*, vol. 379, no. 9834, pp. 2373–2382, 2012.

[36] E. Iob, A. Steptoe, and D. Fancourt, "Abuse, self-harm and suicidal ideation in the U.K. During the COVID-19 pandemic," *Brit. J. Psychiatry*, vol. 217, no. 4, pp. 543–546, Oct. 2020.

[37] M. I. Islam, R. Khanam, and E. Kabir, "Depression and anxiety have a larger impact on bullied girls than on boys to experience self-harm and suicidality: A mediation analysis," *J. Affect. Disorders*, vol. 297, pp. 250–258, Jan. 2022.

[38] S.-P. Jun, H. S. Yoo, and S. Choi, "Ten years of research change using Google trends: From the perspective of big data utilizations and applications," *Technol. Forecasting Social Change*, vol. 130, pp. 69–87, May 2018.

[39] S. Kandula, M. Olfson, M. S. Gould, K. M. Keyes, and J. Shaman, "Hindcasts and forecasts of suicide mortality in U.S.: A modeling study," *PLOS Comput. Biol.*, vol. 19, no. 3, Mar. 2023, Art. no. e1010945.

[40] R. C. Kessler, R. M. Bossarte, A. Luedtke, A. M. Zaslavsky, and J. R. Zubizarreta, "Suicide prediction models: A critical review of recent research with recommendations for the way forward," *Mol. Psychiatry*, vol. 25, no. 1, pp. 168–179, Jan. 2020.

[41] I. Kinchin, C. M. Doran, W. D. Hall, and C. Meurk, "Understanding the true economic impact of self-harming behaviour," *Lancet Psychiatry*, vol. 4, no. 12, pp. 900–901, Dec. 2017.

[42] O. Kraaijeveld and J. De Smedt, "The predictive power of public Twitter sentiment for forecasting cryptocurrency prices," *J. Int. Financial Markets, Institutions Money*, vol. 65, Mar. 2020, Art. no. 101188.

[43] Y.-H. Kuai and H.-L. Ser, "COVID-19 situation in Thailand," *Prog. Microbes Mol. Biol.*, vol. 4, no. 1, pp. 1–8, Dec. 2021.

[44] M. J. Kyron, G. R. Hooke, and A. C. Page, "Prediction and network modelling of self-harm through daily self-report and history of self-injury," *Psychol. Med.*, vol. 51, no. 12, pp. 1992–2002, Sep. 2021.

[45] M. M. Large, "The role of prediction in suicide prevention," *Dialogues Clin. Neurosci.*, vol. 20, no. 3, pp. 197–205, Sep. 2018.

[46] J. Z. Leather, R. C. O'Connor, L. Quinlivan, N. Kapur, S. Campbell, and C. J. Armitage, "Healthcare professionals' implementation of national guidelines with patients who self-harm," *J. Psychiatric Res.*, vol. 130, pp. 405–411, Nov. 2020.

[47] A. Liem, B. Prawira, S. Magdalena, M. J. Siandita, and J. Hudiyana, "Predicting self-harm and suicide ideation during the COVID-19 pandemic in indonesia: A nationwide survey report," *BMC Psychiatry*, vol. 22, no. 1, p. 304, Dec. 2022.

[48] Z. Liu, J. Zhang, and Y. Li, "Towards better time series prediction with model-independent, low-dispersion clusters of contextual subsequence embeddings," *Knowl.-Based Syst.*, vol. 235, Jan. 2022, Art. no. 107641.

[49] W.-Y. Loh, "Classification and regression trees," *Wiley Interdiscipl. Rev., Data Mining Knowl. Discovery*, vol. 1, no. 1, pp. 14–23, 2011.

[50] J. Lohner and N. Konrad, "Deliberate self-harm and suicide attempt in custody: Distinguishing features in male inmates' self-injurious behavior," *Int. J. Law Psychiatry*, vol. 29, no. 5, pp. 370–385, Sep. 2006.

[51] D. E. Losada, F. Crestani, and J. Parapar, "Overview of eRisk 2019 early risk prediction on the Internet," in *Experimental IR Meets Multilinguality, Multimodality, and Interaction*. Lugano, Switzerland: Springer, Sep. 2019, pp. 340–357.

[52] D. E. Losada, F. Crestani, and J. Parapar, "eRisk 2020: Self-harm and depression challenges," in *Proc. 42nd Eur. Conf. IR Res. (ECIR)*, Lisbon, Portugal. Cham, Switzerland: Springer, Apr. 2020, pp. 557–563.

[53] J. Luby and S. Kertz, "Increasing suicide rates in early adolescent girls in the United States and the equalization of sex disparity in suicide: The need to investigate the role of social media," *JAMA Netw. Open*, vol. 2, no. 5, 2019, Art. no. e193916.

[54] D. J. C. MacKay, "Bayesian interpolation," *Neural Comput.*, vol. 4, no. 3, pp. 415–447, May 1992.

[55] A. Malhotra and R. Jindal, "Deep learning techniques for suicide and depression detection from online social media: A scoping review," *Appl. Soft Comput.*, vol. 130, Nov. 2022, Art. no. 109713.

[56] S. M. Thippaiah, M. S. Nanjappa, J. G. Gude, E. Voyiaziakis, S. Patwa, B. Birur, and A. Pandurangi, "Non-suicidal self-injury in developing countries: A review," *Int. J. Social Psychiatry*, vol. 67, no. 5, pp. 472–482, Aug. 2021.

[57] A. Manzoor, "Social media as mirror of society," in *Handbook of Research on Advanced Data Mining Techniques and Applications for Business Intelligence*. Hershey, PA, USA: IGI Global, 2017, pp. 128–141.

[58] J. Mellon and C. Prosser, "Twitter and Facebook are not representative of the general population: Political attitudes and demographics of British social media users," *Res. Politics*, vol. 4, no. 3, Jul. 2017, Art. no. 205316801772000.

[59] A. Memon, S. Sharma, S. Mohite, and S. Jain, "The role of online social networking on deliberate self-harm and suicidality in adolescents: A systematized review of literature," *Indian J. Psychiatry*, vol. 60, no. 4, p. 384, 2018.

[60] L. Mo, "An improved ARIMA method based on hybrid dimension reduction and BP neural network," *Acad. J. Comput. Inf. Sci.*, vol. 5, no. 10, pp. 41–47, 2022.

[61] I. Muneer and R. M. A. Nawab, "Cross-lingual text reuse detection at sentence level for English–Urdu language pair," *Comput. Speech Lang.*, vol. 75, Sep. 2022, Art. no. 101381.

[62] T. Noraset, K. Chatrinan, T. Tawichsri, T. Thaipisutikul, and S. Tuarob, "Language-agnostic deep learning framework for automatic monitoring of population-level mental health from social networks," *J. Biomed. Informat.*, vol. 133, Sep. 2022, Art. no. 104145.

[63] J.-A. Occhipinti, D. Rose, A. Skinner, D. Rock, Y. J. C. Song, A. Prodan, S. Rosenberg, L. Freebairn, C. Vacher, and I. B. Hickie, "Sound decision making in uncertain times: Can systems modelling be useful for informing policy and planning for suicide prevention?" *Int. J. Environ. Res. Public Health*, vol. 19, no. 3, p. 1468, 2022.

[64] P. Padmanathan, H. Bould, L. Winstone, P. Moran, and D. Gunnell, "Social media use, economic recession and income inequality in relation to trends in youth suicide in high-income countries: A time trends analysis," *J. Affect. Disorders*, vol. 275, pp. 58–65, Oct. 2020.

[65] J. Parapar, P. Martín-Rodilla, D. E. Losada, and F. Crestani, "eRisk 2021: Pathological gambling, self-harm and depression challenges," in *Proc. 43rd Eur. Conf. IR Res. (ECIR)*. Cham, Switzerland: Springer, Mar./Apr. 2021, pp. 650–656.

[66] E. Paul and D. Fancourt, "Factors influencing self-harm thoughts and behaviours over the first year of the COVID-19 pandemic in the U.K.: Longitudinal analysis of 49 324 adults," *Brit. J. Psychiatry*, vol. 220, no. 1, pp. 31–37, Jan. 2022.

[67] M. J. Paul, M. Dredze, and D. Broniatowski, "Twitter improves influenza forecasting," *PLoS Currents*, Jun. 2014.

[68] Y. Pei, S. Chen, Z. Ke, W. Silamu, and Q. Guo, "AB-LaBSE: Uyghur sentiment analysis via the pre-training model with BiLSTM," *Appl. Sci.*, vol. 12, no. 3, p. 1182, Jan. 2022.

[69] A. Pourmand, J. Roberson, A. Caggiula, N. Monsalve, M. Rahimi, and V. Torres-Llenza, "Social media and suicide: A review of technology-based epidemiology and risk assessment," *Telemedicine e-Health*, vol. 25, no. 10, pp. 880–888, Oct. 2019.

[70] I. Pramukti, C. Strong, Y. Sitthimongkol, A. Setiawan, M. G. R. Pandin, C.-F. Yen, C.-Y. Lin, M. D. Griffiths, and N.-Y. Ko, "Anxiety and suicidal thoughts during the COVID-19 pandemic: Cross-country comparative study among Indonesian, Taiwanese, and Thai university students," *J. Med. Internet Res.*, vol. 22, no. 12, Dec. 2020, Art. no. e24487.

[71] A. Preti and G. Lentini, "Forecast models for suicide: Time-series analysis with data from Italy," *Chronobiology Int.*, vol. 33, no. 9, pp. 1235–1246, Oct. 2016.

[72] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: Unbiased boosting with categorical features," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–11.

[73] L. Pujante-Otalora, B. Canovas-Segura, M. Campos, and J. M. Juarez, "The use of networks in spatial and temporal computational models for outbreak spread in epidemiology: A systematic review," *J. Biomed. Informat.*, vol. 143, Jul. 2023, Art. no. 104422.

[74] F. Rahman, R. T. Webb, and A. Wittkowski, "Risk factors for self-harm repetition in adolescents: A systematic review," *Clin. Psychol. Rev.*, vol. 88, Aug. 2021, Art. no. 102048.

[75] E. R. Kumar, K. V. S. N. R. Rao, S. R. Nayak, and R. Chandra, "Suicidal ideation prediction in Twitter data using machine learning techniques," *J. Interdiscipl. Math.*, vol. 23, no. 1, pp. 117–125, Jan. 2020.

[76] A. Richardson, T. van Florenstein Mulder, and T. Vehbi, "Nowcasting GDP using machine-learning algorithms: A real-time assessment," *Int. J. Forecasting*, vol. 37, no. 2, pp. 941–948, Apr. 2021.

[77] E. A. Ríssola, M. Aliannejadi, and F. Crestani, "Mental disorders on online social media through the lens of language and behaviour: Analysis and visualisation," *Inf. Process. Manage.*, vol. 59, no. 3, May 2022, Art. no. 102890.

[78] J. Robinson, G. Cox, E. Bailey, S. Hetrick, M. Rodrigues, S. Fisher, and H. Herrman, "Social media and suicide prevention: A systematic review," *Early Intervent Psychiatry*, vol. 10, no. 2, pp. 103–121, Apr. 2016.

[79] S. E. Rodríguez, H. Allende-Cid, and H. Allende, "Detecting hate speech in cross-lingual and multi-lingual settings using language agnostic representations," in *Proc. Iberoamerican Congr. Pattern Recognit.* Cham, Switzerland: Springer, 2021, pp. 77–87.

[80] B. Mahaki, M. Rostami, A. Jalilian, and J. Poorolajal, "Time series analysis of monthly suicide rates in west of Iran, 2006–2013," *Int. J. Preventive Med.*, vol. 10, no. 1, p. 78, 2019.

[81] A. Roy, K. Nikolitch, R. McGinn, S. Jinah, W. Klement, and Z. A. Kaminsky, "A machine learning approach predicts future risk to suicidal ideation from social media data," *npj Digit. Med.*, vol. 3, no. 1, p. 78, May 2020.

[82] S. Scherr, "Social media, self-harm, and suicide," *Current Opinion Psychol.*, vol. 46, Aug. 2022, Art. no. 101311.

[83] H. Schütze, C. D. Manning, and P. Raghavan, *Introduction to Information Retrieval*, vol. 39. Cambridge, U.K.: Cambridge Univ. Press, 2008.

[84] R. Sedgwick, S. Epstein, R. Dutta, and D. Ougrin, "Social media, Internet use and suicide attempts in adolescents," *Current Opinion Psychiatry*, vol. 32, no. 6, pp. 534–541, 2019.

[85] M. R. Segal, "Machine learning benchmarks and random forest regression," Division Biostatist., Univ. California, San Francisco, San Francisco, CA, USA, Tech. Rep., 2004.

[86] H.-C. Shing, S. Nair, A. Zirikly, M. Friedenberg, H. Daumé III, and P. Resnik, "Expert, crowdsourced, and machine assessment of suicide risk via online postings," in *Proc. 5th Workshop Comput. Linguistics Clin. Psychol., From Keyboard Clinic*, 2018, pp. 25–36.

[87] P. Singh, Y. K. Dwivedi, K. S. Kahlon, R. S. Sawhney, A. A. Alalwan, and N. P. Rana, "Smart monitoring and controlling of government policies using social media and cloud computing," *Inf. Syst. Frontiers*, vol. 22, pp. 315–337, Apr. 2019.

[88] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to fine-tune BERT for text classification?" in *Proc. China Nat. Conf. Chin. Comput. Linguistics*. Cham, Switzerland: Springer, 2019, pp. 194–206.

[89] A. Sundar, A. Ramakrishnan, A. Balaji, and T. Durairaj, "Hope speech detection for dravidian languages using cross-lingual embeddings with stacked encoder architecture," *Social Netw. Comput. Sci.*, vol. 3, no. 1, pp. 1–15, Jan. 2022.

[90] P. K. Swain, M. R. Tripathy, S. Priyadarshini, and S. K. Acharya, "Forecasting suicide rates in India: An empirical exposition," *PLoS ONE*, vol. 16, no. 7, Jul. 2021, Art. no. e0255342.

[91] U. S. Tran, R. Andel, T. Niederkrotenthaler, B. Till, V. Ajdacic-Gross, and M. Voracek, "Low validity of Google trends for behavioral forecasting of national suicide rates," *PLoS ONE*, vol. 12, no. 8, Aug. 2017, Art. no. e0183149.

[92] S. Tuarob, C. S. Tucker, S. Kumara, C. L. Giles, A. L. Pincus, D. E. Conroy, and N. Ram, "How are you feeling?: A personalized methodology for predicting mental states from temporally observable physical and behavioral information," *J. Biomed. Informat.*, vol. 68, pp. 1–19, Apr. 2017.

[93] J. M. Twenge, "Increases in depression, self-harm, and suicide among U.S. Adolescents after 2012 and links to technology use: Possible mechanisms," *Psychiatric Res. Clin. Pract.*, vol. 2, no. 1, pp. 19–25, Jun. 2020.

[94] G. Vigderhous, "Forecasting sociological phenomena: Application of Box-Jenkins methodology to suicide rates," *Sociol. Methodol.*, vol. 9, pp. 20–51, Jan. 1978.

[95] Y. Wang, J. Tang, J. Li, B. Li, Y. Wan, C. Mellina, N. O'Hare, and Y. Chang, "Understanding and discovering deliberate self-harm content in social media," in *Proc. 26th Int. Conf. World Wide Web*, Apr. 2017, pp. 93–102.

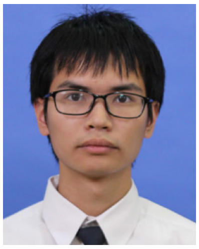
[96] R. Yorsaeng, N. Suntronwong, I. Thongpan, W. Chuchaona, F. B. Lestari, S. Pasittungkul, J. Puenpa, K. Atsawawaranunt, C. Sharma, N. Sudhinaraset, A. Mungaomklang, R. Kitphati, N. Wanlapakorn, and Y. Poovorawan, "The impact of COVID-19 and control measures on public health in Thailand, 2020," *PeerJ*, vol. 10, Feb. 2022, Art. no. e12960.

[97] J. Yu et al., "Seasonality of suicide: A multi-country multi-community observational study," *Epidemiol. Psychiatric Sci.*, vol. 29, p. e163, Aug. 2020.

[98] G. Zerveas, S. Jayaraman, D. Patel, A. Bhamidipaty, and C. Eickhoff, "A transformer-based framework for multivariate time series representation learning," in *Proc. 27th ACM SIGKDD Conf. Knowl. Discovery Data Mining*, Aug. 2021, pp. 2114–2124.

[99] A. Zirikly, P. Resnik, O. Uzuner, and K. Hollingshead, "CLPsych 2019 shared task: Predicting the degree of suicide risk in Reddit posts," in *Proc. 6th Workshop Comput. Linguistics Clin. Psychol.*, 2019, pp. 24–33.

SUPPAWONG TUAROB (Member, IEEE) received the B.S.E. and M.S.E. degrees in computer science and engineering from the University of Michigan-Ann Arbor and the M.S. degree in industrial engineering and the Ph.D. degree in computer science and engineering from the Pennsylvania State University. Currently, he is an Associate Professor of computer science at the Faculty of Information and Communication Technology, Mahidol University, Thailand. His research interests include data mining in large-scale scholarly, social media, and healthcare domains, as well as applications of intelligent technologies for social good.



KRITTIN CHATRINAN received the bachelor's degree from the Faculty of Information and Communication Technology, Mahidol University, Thailand, where he is currently pursuing the master's degree in computer science. His research interests include natural language processing and applying machine learning methods in social media and healthcare fields.

THANAPON NORASET received the B.Sc. degree from Mahidol University, Thailand, in 2007, and the Ph.D. degree in computer science from Northwestern University, USA, in 2017. He is currently a Faculty Member at the Faculty of Information and Communication Technology, Mahidol University. His research interests include natural language processing and machine learning.

TANISA TAWICHSRI received the bachelor's degree (magna cum laude) in mathematics and economics from Washington University in St. Louis, the master's degrees in economics and statistics from Arizona State University, and the Ph.D. degree in economics from Arizona State University, in 2018. Currently, she serves as a Researcher at the Puey Ungphakorn Institute for Economic Research (PIER). Before joining PIER in June 2019, she worked at the National Accounts Office and the Office of the National Economics and Social Development Council. Her research interests include the public economics and applied microeconomics. She has worked on topics such as smoking bans, tax policies, and household consumption responses. Her current research interests include labor economics and health economics.

TIPAJIN THAIPISUTIKUL received the master's degree (Hons.) in research path from The University of Sydney (USYD), Sydney, NSW, Australia, in 2012, and the Ph.D. degree from the Department of Computer Science and Information Engineering, National Central University, Taiwan, in 2021. She is currently an Instructor with the Faculty of Information and Communication Technology (ICT), Mahidol University, Salaya, Thailand. Her research interests include machine learning, applied intelligence, data mining, and social network analysis.