Journal of the American Medical Informatics Association, 30(6), 2023, 1068–1078
https://doi.org/10.1093/jamia/ocad062
Advance Access Publication Date: 12 April 2023
Research and Applications

Research and Applications

From benchmark to bedside: transfer learning from social media to patient-provider text messages for suicide risk prediction

Hannah A. Burkhardt [1], Xiruo Ding[1], Amanda Kerbrat[2], Katherine Anne Comtois[2], and Trevor Cohen[1]

[1]Department of Biomedical Informatics and Medical Education, University of Washington, Seattle, Washington, USA and [2]Department of Psychiatry and Behavioral Sciences, University of Washington, Seattle, Washington, USA

Corresponding Author: Hannah A. Burkhardt, Biomedical Informatics and Medical Education, University of Washington, 850 Republican St, Seattle, WA 98109, USA; haalbu@uw.edu

Received 12 December 2022; Revised 6 March 2023; Editorial Decision 25 March 2023; Accepted 28 March 2023

ABSTRACT

Objective: Compared to natural language processing research investigating suicide risk prediction with social media (SM) data, research utilizing data from clinical settings are scarce. However, the utility of models trained on SM data in text from clinical settings remains unclear. In addition, commonly used performance metrics do not directly translate to operational value in a real-world deployment. The objectives of this study were to evaluate the utility of SM-derived training data for suicide risk prediction in a clinical setting and to develop a metric of the clinical utility of automated triage of patient messages for suicide risk.

Materials and Methods: Using clinical data, we developed a Bidirectional Encoder Representations from Transformers-based suicide risk detection model to identify messages indicating potential suicide risk. We used both annotated and unlabeled suicide-related SM posts for multi-stage transfer learning, leveraging customized contemporary learning rate schedules. We also developed a novel metric estimating predictive models' potential to reduce follow-up delays with patients in distress and used it to assess model utility.

Results: Multi-stage transfer learning from SM data outperformed baseline approaches by traditional classification performance metrics, improving performance from 0.734 to a best F1 score of 0.797. Using this approach for automated triage could reduce response times by 15 minutes per urgent message.

Discussion: Despite differences in data characteristics and distribution, publicly available SM data benefit clinical suicide risk prediction when used in conjunction with contemporary transfer learning techniques. Estimates of time saved due to automated triage indicate the potential for the practical impact of such models when deployed as part of established suicide prevention interventions.

Conclusions: This work demonstrates a pathway for leveraging publicly available SM data toward improving risk assessment, paving the way for better clinical care and improved clinical outcomes.

Key words: social media, natural language processing, artificial intelligence, suicide prevention, decision-making, computer assisted, delivery of health care

INTRODUCTION

Suicide is now the tenth leading cause of death in the United States.[1] Regular suicide risk screening at primary care encounters is effective in identifying individuals at elevated risk,[2] and the Joint Commission now prescribes several suicide prevention measures for healthcare organizations.[3,4] Current evidence supports the efficacy of postdischarge follow-up contacts via interventions such as Caring Contacts,[5,6] which entails suicide prevention professionals periodically sending brief messages of unconditional care and concern to individuals in need of support. This intervention reduces suicidal thoughts and behaviors, suicide attempts, and suicide completion.[7] Programs tend to enroll individuals with known risk, for example, those with a suicide-related encounter, and have been conducted via postal mail, email, and text message. In intervention designs with two-way communication, patients may reply to messages if they wish and will receive further tailored support from intervention staff in response.

Although two-way communication provides additional opportunities for effective support, this intervention design imposes additional burdens for risk management: if participants reach out with an urgent need for help, timely and effective follow-up is critical. Therefore, resource shortages have precluded the broad adoption of this intervention. In our prior work, we conducted extensive interviews with clinical and administrative stakeholders experienced with Caring Contacts interventions, revealing opportunities for machine learning (ML) methods to alleviate this workload burden and provide scalable suicide crisis support. Specifically, natural language messages sent by patients to intervention staff may be automatically triaged for urgent follow-up.[8]

Data challenges in clinical settings

However, the development of predictive models is hampered by the size and availability of clinical suicide risk datasets. Training state-of-the-art neural network-based natural language processing (NLP) models from scratch requires considerable amounts of data.[9] In recent years, neural transfer learning approaches leveraging unlabeled datasets have alleviated this requirement,[10] but models still perform best with thousands of labeled examples or more. Obtaining this amount of labeled data is typically not feasible in clinical settings, where data privacy and security are carefully controlled,[11,12] making it difficult to share labeled datasets for reuse or to collaborate on benchmark dataset development. Resource limitations in healthcare settings may also impede annotation efforts locally. In addition, while existing real-world healthcare data, such as from EHRs, can be harvested for secondary use, patient-generated natural language data is not commonly collected as part of routine healthcare operations, so those data must be purposely collected, further limiting the size of related datasets.

In contrast, vast amounts of publicly available social media (SM) data are continuously generated. The comparatively lower level of privacy concern also allows outsourcing of annotation efforts, for example, to crowdsourcing platforms. Suicidal individuals reach out to peers in informal settings such as on SM platforms, and a substantial amount of published work investigates detecting suicide risk from SM posts. For example, Coppersmith et al[13] predicted expert suicide risk assessments from a range of SM data, and Shing et al[14] created a publicly available annotated dataset of Reddit posts and demonstrated the ability to assess suicide risk in this benchmark dataset.[14] In the 2019 CLPsych Shared Task challenge, several ML approaches were benchmarked in this dataset, with deep learning techniques performing best.[15] The clinical and SM settings differ in many respects, but there is also substantial overlap. If datasets, trained models, and findings from SM research could be effectively transferred to clinical risk prediction, the data challenges in this setting would be alleviated.

Transfer learning

Prior work suggests that information is transferable between domains and tasks.[16] Related but distinct data, for example, with a different input feature space or data distribution,[17] can augment learning for another task with data scarcity. Selecting a domain-specific pretrained language model is a well-established approach to leveraging multiple data sources to boost performance.[18,19] After pretraining on general-purpose text, such models are further pretrained on domain-specific corpora, for example on biomedical research literature, for improved performance on biomedical tasks.[20] In addition, domain-adapted pretrained models must be fine-tuned for individual prediction tasks using labeled data. Multi-stage pretraining and fine-tuning protocols have been proposed, notably by Howard and Ruder.[21] Here, we evaluate the hypothesis that neural transfer learning from larger, publicly available SM datasets can mitigate the problem of small dataset size in clinical suicide risk prediction.

Measuring clinical utility

While artificial intelligence (AI) has the potential to fundamentally transform the way healthcare is delivered, a lack of demonstrated operational utility has been identified as a significant barrier to achieving the adoption of AI into clinical practice.[22] Clinical trials represent the gold standard of evidence in biomedicine (including for biomedical AI[23]) to establish the safety and efficacy of drugs, interventions, and medical devices, but the desire to accelerate adoption is difficult to align with their substantial time and resource requirements.[24] Accordingly, model utility should be established before proceeding with clinical trials. In contrast, standardized performance metrics, such as the area under the receiver operator characteristic curve (AUROC), are easy to assess, widely used, and generally recommended in ML research. While such metrics are indispensable to making model performance comparable across settings, the ultimate value of a predictive model intended for clinical use is determined by its real-world impact on healthcare delivery, which depends on many other factors also. Evaluation approaches should therefore consider these factors by incorporating parameters of the deployment environment that can be assessed at development time. For example, Jung et al[25] evaluated the clinical utility of a model recommending advanced care planning and found that the model's utility was constrained by the health system's capacity to provide this service. Their evaluation informed a strategy for harmonizing the model with practical constraints to realize optimal utility. Bayati et al[26] performed a cost-effectiveness analysis of the impact of intervening based on model predictions, incorporating both the cost of a postdischarge intervention and its potential to reduce near-term readmission. Here, we develop a utility metric, average time to response in urgent messages (ATRIUM), for a suicide risk prediction model within the Caring Contacts workflow, and use it to evaluate our models.

MATERIALS AND METHODS

Data

We used the University of Maryland Reddit Suicidality Dataset (Version 2), a dataset of Reddit posts from suicidal individuals curated by Shing et al,[14,15] henceforth the SM dataset. Specifically, we used the portion of the dataset annotated for the task of flagging

Table 1. Data characteristics

	SM	CCVT
Context		
Setting	Social media platform Reddit	Clinical trial among military service members
Population	Reddit users proactively seeking advice online	Help-seeking individuals identified by clinical experts to be at suicide risk
Message audience/confidentiality	Anonymous posts to the general public	Initiated in a confidential one-on-one setting with a suicide prevention professional; text messages from/to personal mobile phones
Purpose of messages	Social media posts, for example, to seek advice or empathy	Caring Contacts
Time frame	2008–2015	2013–2017
Participant characteristics		
Number of participants	621	221
Sex male	69%[a]	82%
Age	58% 18–29[a] 33% 30–49[a] 7% 50–64[a] 1% 65+[a]	Mean 25.6 (SD 6.3)
Race/ethnicity		
• White	63%[a]	66.0%
• Black	10%[a]	10.0%
• Hispanic	14%[a]	9.1%
Message characteristics		
Message count	1105	1229
Indicating risk/urgency	82.3%	18.2%
Words in message, mean (SD)	222.0 (250.9)	9.3 (10.2)
Messages per user		
Mean (SD)	1.8 (2.4)	5.6 (4.9)
25th percentile	1.0	2.0
Median	1.0	4.0
75th percentile	2.0	8.0

Abbreviations: CCVT: Caring Contacts Via Text; SM: social media.
[a]Pew Research Center estimates.[27]

high-risk individuals amongst Redditors with posts in the/r/SuicideWatch subreddit. In the study described by Shing et al,[14] user risk levels were rated by crowd workers and four experts: a suicide prevention coordinator at the Veteran's Administration, a committee cochair at National Suicide Prevention Lifelines Standards, a doctoral student with training in suicide assessment and treatment, and an ED psychiatrist. Posts were rated on the following scale: (1) no risk, (2) low risk, (3) moderate risk, and (4) severe risk. All posts belonging to the same user receive the same class label, which is the highest label across each user's posts. The "flagging" task consists of distinguishing unconcerning utterances (1) from those indicating an elevated risk that may warrant intervention (2–4). Dataset characteristics are shown in Table 1. For additional details on the data collection and labeling process, the reader is referred to Shing et al.[14]

The clinical dataset contains text messages from the Caring Contacts Via Text (CCVT) clinical trial, which investigated the effectiveness of Caring Contacts to reduce suicidal thoughts and behaviors.[5] In this trial, Comtois et al randomized military service members across three military installations to either usual care or a text message-based Caring Contacts intervention combined with usual care. All participants reported suicidal ideation at baseline, and 44.3% had previously attempted suicide. Participants in the Caring Contacts condition received 11 text messages during the 12-month intervention period.[5] When CCVT patients responded, study staff reciprocated based on their clinical judgment of whether the patient needed urgent support and availability. A patient may respond to an automatically scheduled message, for example, "Hi there, hope your week has been good!", with a message indicating distress, for example, "Thanks, but this week has been horrible." While the study protocol did not prescribe any particular response time frame, study staff treated such messages more urgently than others. If the message indicated an acute crisis, staff immediately followed up with a text message and a phone call, rather than a text message alone. In contrast, a patient may respond positively or neutrally, for example, "Thank you!". Most messages of this kind were still responded to, but without urgency, that is, with some delay. Staff were notified of every incoming message, and judged urgency as messages were received. After the trial concluded, study staff annotated responses for the level of distress expressed by the message author (ie, message urgency) on the following scale: 0—none; 1—difficulty; 2—nonurgent distress; and 3—urgent distress/crisis. Dataset characteristics are shown in Table 1.

The SM dataset contains over 20 times more total words than the clinical dataset. Neural language models can learn from every word contained in the training data, that is, parameters relating to every word in the input sequence are updated. Therefore, although the number of examples is similar, the SM dataset contains a substantial amount of additional information.

The datasets have skewed but similar message frequency distributions across users: In the CCVT data set, the top 10%, 20%, and

50% of users account for 29.8%, 48.5%, and 82.0% of the messages, respectively. In the SM data set, the top 10%, 20%, and 50% of users account for 38.7%, 51.0%, and 71.9% of the messages, respectively.

The two datasets were aligned for the single binary classification task of flagging messages for follow-up, that is, distinguishing messages labeled "a" from those labeled "b", "c", or "d", and messages labeled "0" from those labeled "1", "2", or "3", respectively. This aligns with the intended purpose of the SM "flagging" data subset.

Distribution and data characteristics differ between these data sets (Table 1). Texts were produced by different people, in different settings, for different purposes, and they differ in length and class distribution. These differences pose challenges for transfer learning. Serendipitously, the two datasets also have significant similarities. Both of our datasets contain text written by an individual predisposed to suicidal thoughts and behaviors, population demographics are similar and even though label definitions have different nuances, the labels can be grouped in the same way (ie, no risk vs any level of risk). Beyond these domain and task-specific similarities, prior work shows that the language in text messages and SM posts is generally comparable in terms of word frequency, lexical density, and emotional expressions, although SM language may be more formal.[28,29]

Bidirectional Encoder Representations from Transformers model

Bidirectional Encoder Representations from Transformers (BERT) is a deep learning architecture for NLP first reported by Devlin and colleagues in 2018.[30] BERT was initially trained on large amounts of natural language in an unsupervised (or "semi-supervised") manner, learning to predict held-out ("masked") words and the sequence in which sentences occur. This pretraining informs the contextual representations it derives from previously unseen text, providing both initial representations of words (or their components) and initial weights through which to estimate their influence in context. These contextual representations can then be used for various prediction tasks, including text classification.

Further pretraining on a domain-relevant corpus is recommended.[10] Several domain-specific pretrained BERT models are available; for example, BioBERT, trained on biomedical research literature, improves upon base BERT in biomedical tasks.[20] Here, we used Public Health Surveillance (PHS)-BERT[18] as a base model because it has a demonstrated performance advantage on suicide-related prediction tasks compared to other biomedically relevant pretrained models such as BioBERT and the mental health-specific MentalBERT.[19] In addition, in preliminary exploratory work, PHS-BERT outperformed other BERT models in our setting.

Utility metric: average time to response in urgent messages

We present a way to measure clinical utility in terms of the average time to response in urgent messages (ATRIUM). The metric assumes a limited work capacity to respond to messages quickly (eg, three messages per hour), with all remaining messages being deprioritized (eg, responded to at the end of the day). The algorithmic problem is to assign priority such that as many urgent messages as possible are prioritized. Accordingly, the triage algorithm should assign higher scores to urgent messages, with the top-ranked messages addressed immediately. Assuming a work capacity k, model performance might be measured in terms of precision at k. However, we are also interested in minimizing the number of erroneously deprioritized urgent messages. We therefore define a, the number of urgent messages within the top k; and U, the total number of urgent messages, with a/k equivalent to the precision at k; the number of missed urgent messages is $U - a$. U depends on the data and is therefore treated as a constant; a is specific to the model; k, in practice, is a tradeoff between costs and benefits, that is, staff availability compared to the amount of time that can be saved. It will depend on the organization.

Let T_{urgent} and $T_{\text{nonurgent}}$ be the response times that can be expected for prioritized and deprioritized messages, respectively. ATRIUM is the average response time across urgent messages, given that a of them will correctly be treated with urgency, that is, responded to with a delay of T_{urgent}, and the remaining $U - a$ messages will incorrectly be treated without urgency, that is, responded to with a delay of $T_{\text{non-urgent}}$. In other words, ATRIUM is a weighted average of quick and slow response times, as follows:

$$\text{ATRIUM}(a,\ U) = \left(\frac{a}{U}\right) T_{\text{urgent}} + \left(1 - \frac{a}{U}\right) T_{\text{nonurgent}}$$

We assume that in our dataset, all urgent messages were treated with priority, and all nonurgent messages were deprioritized, that is, $k = a = U$. Therefore, we drew upon the observed response times in the CCVT clinical study dataset for messages annotated as urgent and nonurgent to calculate T_{urgent} and $T_{\text{nonurgent}}$, capturing what might be expected for a message treated as urgent compared to one treated as nonurgent.

We use all messages where paired timestamps were available, that is, where a patient replied and staff responded to the reply with another text message. Some responses were sent after more than 600 minutes; manual review revealed reasons of technical error (message did not go through) and responsible staff being out of the office (response sent between 8 am and 9 am the following day). For this analysis, we consider these to be nonrepresentative of the response time distribution and therefore exclude them. A total of 421 patient messages with timestamped responses remained, with 79 (18.8%) expressing some level of difficulty, and 321 (81.2%) being positive or neutral. Response times were approximately normally distributed for each class, with a slight skew toward earlier times for urgent messages. The average response times were $T_{\text{urgent}} = 106.2$ minutes and $T_{\text{nonurgent}} = 130.6$ minutes, respectively.

Baselines

The upper and lower bounds for performance are the random assignment of priorities and human triage, respectively, as described in Table 2.

Calculation of expected values for ATRIUM

To calculate the expected value of ATRIUM for each model, we first determine a at each $k \in \{1, \ldots, N\}$ where N is the total number of messages in the dataset. This is accomplished as described in Table 2 for the human and random baselines; for the ML models, we rank order the set of 421 predictions corresponding to timestamped messages, and count the number of correct instances in the top-ranked k predictions, for each k. Next, we determine the expected values for ATRIUM for each model, given U and a for each k, using anticipated urgent and nonurgent response times from previous or expert knowledge. Here, we use $T_{\text{urgent}} = 106.2$ minutes and $T_{\text{nonurgent}} = 130.6$ minutes.

Model development

We trained and evaluated a bag-of-words baseline model, a BERT-based suicide risk classifier, and a BERT-based suicide risk prediction classifier that additionally leverages supervised transfer learning from an annotated SM dataset (Figure 1).

Table 2. Definition of random and human baselines

Baseline	Formula	Description
Random baseline	$a = k\ \left(\frac{U}{N}\right)$	• Messages are ranked randomly • a follows a hypergeometric probability distribution (the probability of a successes in k draws, without replacement, from a finite population of size N that contains exactly U objects of interest) • To be improved upon by any alternative approach
Human baseline	$a = \begin{cases} U\ if\ k > U \\ k\ otherwise \end{cases}$	• The CCVT trial was staffed such that staff shortages would not be responsible for patients in urgent need receiving a delayed response, that is, every urgent message was recognized as urgent and handled accordingly (recognizing that response times will vary due to many real-life factors) • In this dataset, there were 79 urgent messages • Thus, we assume that the CCVT dataset reflects $k = a = U = 79$

Note: a: the number of correctly predicted urgent messages among the top k predictions. k: work capacity for treating messages with priority; depends on staffing. U: total number of urgent messages.

Abbreviation: CCVT: Caring Contacts Via Text.

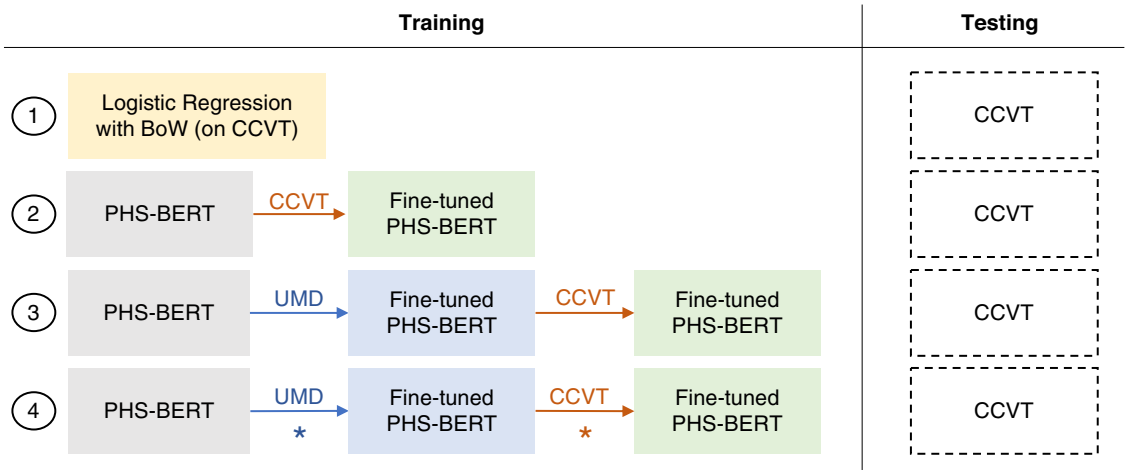

Figure 1. Model and data flow overview. * marks where the Howard and Ruder transfer learning technique was used.

Training and performance calculation using cross-validation

Models were evaluated using 5-fold cross-validation, splitting data into five subsamples in a stratified manner by user, such that all documents from a single participant appeared in the same fold and the label frequency was approximately retained within each fold. In each of five rounds, four folds of the data were pooled and used to train the model as described below, and predictions were produced for the remaining fold. We trained five models in this way to obtain one prediction for each instance in the dataset. Performance metrics were calculated on this set of predictions, resulting in one score per model. To account for the possibility of a nonrepresentative randomized split, this process was repeated for a total of five different random splits of the data, and the median of the five resulting scores is reported. For ATRIUM, we ranked only the predictions for messages for which paired timestamps were available to calculate a and calculate the ATRIUM on this subset.

Bag-of-words baseline

We removed stop words, punctuation, and special characters from each document and lemmatized words using the Natural Language Tool Kit (NLTK).[31] Then, for each document j, we constructed a feature vector $b_j = [w_0,\ w_1,\ \ldots w_n]$ where n is the number of unique words in the corpus (here, 1220) and w_i is the count of that word in document j. We then fit logistic regression, support vector machine, and random forest models and determined performance across the entire set using the cross-validation data splits and training procedure described previously. Hyperparameters were selected with nested cross-validation[32] using Scikit-Learn's GridSearchCV,[33] applied to a validation partition (10%) of each training split. Preliminary work also investigated term frequency-inverse document frequency (TF-IDF) transformations; we do not report results with TF-IDF weighting, as this led to a drop in baseline model performance, possibly due to the corpus being too small to provide helpful estimates of lexical frequency.

BERT models

All BERT models were trained on 90% of each training split for up to 12 epochs, with the best-performing epoch selected according to the cross-entropy loss calculated on the remaining 10% of the training data (the validation set). To account for class imbalance, in all BERT models, the loss function (cross-entropy) was modified to weight classes according to the reciprocal of their frequencies in the training data, that is, approximately 5:1 and 1:5 for the SM and CCVT datasets, respectively.

The first BERT model was trained by fine-tuning PHS-BERT (from the Huggingface library of pretrained models[34]) on the CCVT dataset.

The second BERT model was trained in two stages. We first fine-tuned PHS-BERT with the SM dataset and then continued the training process using the CCVT data.

We additionally trained a BERT model using a contemporary approach to selecting learning rates to optimize multi-stage transfer learning described by Howard and Ruder.[21] The three techniques involved—discriminative fine-tuning, slanted triangular learning rates (STLR), and gradual unfreezing—are aimed at counteracting the loss of knowledge from prior pretraining due to overly aggressive fine-tuning, also known as catastrophic forgetting. Discriminative fine-tuning involves selecting learning rates that are exponentially larger for later layers than earlier ones. Tuning each layer with a different learning rate allows more prior knowledge to be retained in earlier layers. With STLR, learning rates are linearly increased for a set number of training iterations, then linearly decreased for the remaining iterations, allowing the model to first select a general region of the parameter space, and then slowly converge within that region. Gradual unfreezing refers to first freezing the layers of the model, that is, setting the learning rates to zero, and then unfreezing one layer per epoch, starting with the last layer. For a detailed description of these techniques, see Howard and Ruder.[21]

In our first transfer learning phase, we used the first two techniques for fine-tuning the PHS-BERT model using the SM dataset. We combined all three techniques in the next transfer learning phase to further tune the model to the CCVT data.

RESULTS

Our experiments were designed to answer two main questions. The first was whether transfer learning from the SM set would improve performance with our clinically derived data. As shown in Table 3, all models achieve high AUROCs, with precision and recall well balanced in all neural models, possibly due to class weighting. In terms of F1 score, at 0.797, the best transfer learning approach achieved a 6.3% (0.063) improvement over the BERT model that did not use transfer learning (F1 0.734), and a 21.2% (0.212) improvement over the best bag-of-words baseline model (F1 0.585). The best transfer learning approach, which used learning rate scheduling optimized for transfer learning proposed by Howard and Ruder, improved upon the baseline transfer learning model by 1.3% (0.013). These are substantial improvements in performance for deep learning models over classical ML approaches, for the BERT models with transfer learning over their counterpart trained without this step, and for the model with customized learning rate schedules over all others.

The second question concerned the extent to which a measure of clinical utility would reflect improvements in classification performance. In Table 4 and Figure 2, we present the estimates for the time saved on average per urgent message due to the use of each model as a better (ie, more relevant to a clinical setting) approximation of model performance. The difference between ATRIUM in a scenario without triage (random baseline) and the ATRIUM achievable with predictive model use is shown in parentheses in Table 4.

The random baseline model represents the time to response expected without triage. The bag-of-words model markedly improves upon this random baseline, saving approximately 10.4 minutes for each urgent message with $k = 100$. The nontransfer BERT model outperforms the bag-of-words baseline, saving an estimated 14.1 minutes per urgent message. The best-performing model leverages transfer learning with Howard and Ruder's strategy for selecting learning rates, saving about 15.9 minutes per urgent message over the no-triage baseline. Response times for urgent messages that might be expected when using this classifier to perform automatic triage closely resemble what might be expected with human triage: approximately 18.9 minutes saved per urgent message over random triage at $k = 100$. These estimates show that the advantages in classification performance shown in Table 3 indicate faster response times, of a magnitude that suggests the potential for real clinical impact.

As expected, more time can be saved with k set higher; this is true even for the random triage model. The time saved over random triage therefore begins to decline at the (artificial) optimum with $k = 79$ for the human triage model. The ML models do not have this artificial peak, so this pattern does not hold for them; this is an artifact of the data used in our approach to creating the human baseline, which has 79 positive examples.

DISCUSSION

Our findings show that transfer learning from annotated SM data developed to serve as a benchmark for a shared task improves suicide risk prioritization of messages collected in the context of a clinical intervention. Also, our evaluation approach estimates clinical utility as time saved, which may be more intuitive to clinical stakeholders than conventional metrics.

Transfer learning improved performance substantially. We used both a domain-specific pretrained BERT model with demonstrated success for suicide risk prediction,[18,34] and a dataset of labeled SM

Table 3. Median performance metrics across 5 runs with different cross-validation splits, calculated on aggregated predictions on the test splits

	Acc.	F1	Precision	Recall	AUROC	AUPRC	Precision@79
Bag of words							
Logistic regression	0.868	0.585	0.687	0.509	0.858	0.660	0.595
SVM	0.854	0.581	0.607	0.558	0.801	0.542	0.595
Random forest	0.858	0.455	0.753	0.326	0.890	0.656	0.608
PHS-BERT + CCVT	0.902	0.734	0.721	0.741	0.947	0.825	0.722
PHS-BERT + SM + CCVT	0.924	0.784	**0.798**	0.768	0.955	0.860	**0.785**
PHS-BERT + SM + CCVT with Howard and Ruder LRs	**0.925**	**0.797**	0.793	**0.772**	**0.961**	**0.875**	**0.785**

Note: Precision@79 represents an idealized scenario in which staff are available to triage exactly the number of urgent messages in the set. The best performance for each metric is shown in bold.

Abbreviations: BERT: Bidirectional Encoder Representations from Transformers; CCVT: Caring Contacts Via Text; PHS: Public Health Surveillance; SM: social media; LRs: learning rates.

Table 4. Average time to response in urgent messages (ATRIUM) (time saved compared to baseline), calculated using the model with the median performance metrics

k	20	50	79	100	120	150
0. Random baseline	129.6 (0.0)	127.8 (0.0)	126.1 (0.0)	125.0 (0.0)	123.8 (0.0)	121.7 (0.0)
1. Bag of words (LR)	125.1 (4.5)	119.3 (8.5)	116.2 (10.0)	114.6 (10.4)	112.4 (11.4)	111.6 (10.1)
2. PHS-BERT + CCVT	124.4 (5.1)	117.0 (10.7)	113.1 (13.0)	111.0 (14.1)	109.1 (14.8)	108.2 (13.5)
3. PHS-BERT + SM + CCVT	124.4 (5.2)	116.8 (11.0)	111.6 (14.5)	109.5 (15.6)	107.5 (16.4)	106.9 (14.8)
4. PHS-BERT + SM + CCVT with Howard and Ruder LR	124.3 (5.3)	116.7 (11.1)	111.5 (14.6)	109.1 (15.9)	107.6 (16.2)	107.5 (14.2)
5. Human triage	124.4 (5.2)	115.4 (12.4)	106.2 (19.9)	106.1 (18.9)	106.2 (17.6)	106.2 (15.4)

Note: *k* of 79 represents an idealized scenario in which staff are available to triage exactly the number of urgent messages in the set.

Abbreviations: BERT: Bidirectional Encoder Representations from Transformers; CCVT: Caring Contacts Via Text; PHS: Public Health Surveillance; SM: social media.

posts for additional transfer learning. Though there are similarities, the distribution and characteristics of the SM data differ from the clinical data that is the target of our final model. We found that incorporating this SM data markedly boosts the F1 score, from 0.734 to 0.797 (the latter with customized learning rate schedules). These findings suggest that suicide risk prediction in clinical populations can be improved by leveraging novel transfer learning approaches and publicly available annotated text data.

We also developed a utility metric for models employed for message triage within the Caring Contacts workflow. When evaluating the utility of models, it is important to assess not only standardized performance metrics (AUROC, AUPRC), but also the performance metrics most relevant for the problem setting (eg, precision at *k*). To move further along the continuum from "in-situ" model performance to clinical utility, it is important to assess the model's practical value in quantities relevant and intuitively meaningful to stakeholders. Though substantial, it is not obvious what practical differences the improvements in F1 score would make for the model user. Assessing precision at *k* offers additional insight into clinically relevant variables such as the likely number of false positive alerts that may occur at the point of deployment. Yet, precision at *k* does not capture the magnitude of the effect on the workflow that can be expected. Here, we estimate this clinical impact in terms of time saved per urgent message in minutes. Compared to employing no triage, on average, approximately 16.6 minutes could be shaved off each urgent message response time if this model were to be used for automatic triage of incoming messages (at $k = 79$). This is within approximately 5 minutes per message of the expected time saved with human triage.

This metric can also inform the allocation of staff when scaling the intervention. Scaling allows organizations to benefit more patients, but may incur prohibitive staffing requirements with manual triage. Assessing ATRIUM for each triage approach across a range of values for the work capacity *k* allows comparisons of staffing requirements for each triage approach, given an acceptable average response time. For example, if an organization were to determine that an ATRIUM of 110 min should be targeted, the bag-of-words model would require enough staff to reach a work capacity of almost 170 messages; in contrast, the best BERT model might accomplish an ATRIUM of 110 min at a work capacity of only k=90, corresponding to a 47% reduction in staffing needs. Thus, automated language models can support intervention scalability.

Ethical and practical considerations for suicide prevention on SM versus healthcare settings

Our lives are increasingly digitized, with immense amounts of data automatically created or collected by consumer technologies. Patient-generated data, for example, smartphone logs and SM data, have enabled a proliferation of consumer technologies promising to benefit human health. SM data, in particular, hold unprecedented promise; they have mediated the development of new markers of mental health symptomatology in depression and anxiety.[35–38] Facebook has operationalized ML to detect users with high suicide risk on their platform.[39,40]

These data also represent a unique opportunity to benefit clinical care: recent work, including our own, has demonstrated the potential of utilizing patient-generated natural language data in clinical settings, showing that language indicators extracted from logs of message-based clinical psychotherapy sessions can predict patient trajectories in depression.[41,42]

Indeed, leveraging these data for clinical purposes may be more ethical, practical, and acceptable to users. Barnett and Torous argue that clinicians, not advertising companies, should have the responsibility to identify and help suicidal individuals.[43] Besides the ethical and privacy concerns surrounding publicly traded companies inferring sensitive health information about their users[44,45] and acting on it without explicit consent,[13] commercial entities operating such technologies may not be able to respond appropriately to an emergent crisis.[43] Some have warned that bad actors may purposely target individuals identified as vulnerable in public forums[44,46]; others have argued that only medical professionals should engage in triage and intervention.[43,45] Model developers at Facebook humbly state, "Our expertise at building social networking and scalable software systems in no way qualified us to reinvent suicide prevention".[39] Undoubtedly, trained clinical professionals are best equipped for suicide prevention.

Integrating emergent signals into the clinical workflow to support diagnosis and treatment by qualified clinicians is the most responsible and effective path forward. Yet, the intersection between SM research and clinical decision support research remains small, with a lack of work investigating how to translate the advances made with consumer technology to clinical settings.

The current research advances our understanding of how best to leverage mental health insights originating outside the healthcare setting as part of ongoing clinical care. We demonstrated an approach to applying the advances made with SM data toward clinical risk assessment, empowering qualified clinicians to better evaluate and care for patients. Thus, this research helps realize the translational potential of SM-derived signal to detect suicide risk in the context of an established clinical relationship.

Limitations

The CCVT dataset is comparatively small and specific to the Caring Contacts intervention. Nonetheless, it reflects a real clinical use case

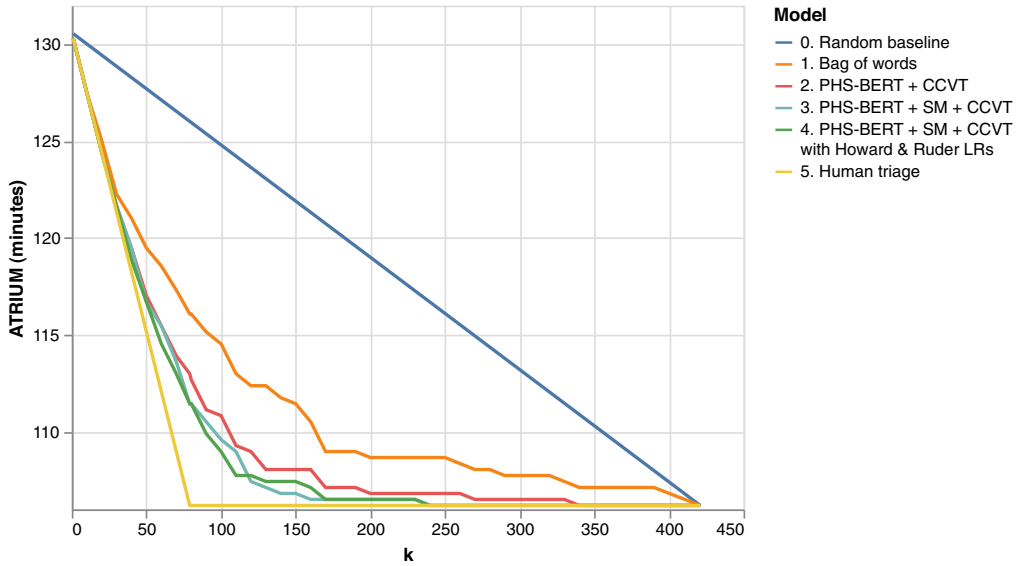

Figure 2. Average time to response in urgent messages (ATRIUM) versus *k*. "Bag of words" is the best bag-of-words model by F1 score, which used logistic regression.

for suicide risk prediction. Different sites will likely have to develop models customized to their Caring Contacts participants; however, our approach demonstrates the feasibility of using larger, more readily available datasets to optimize performance on these smaller ones.

The unique characteristics inherent in SM datasets may be potentially limiting for clinical settings. In our case, the texts in the two datasets were not written by the same set of individuals, so our datasets are not only from different settings, but also from different populations. Here, age, gender, and race were comparable between the populations, which may be one explanation for the substantive improvements in performance demonstrated with transfer learning in this case. However, it is likely that participants have different cultural and socioeconomic backgrounds that affect language use, how they interact with others, and how they use technology. According to a recent review, general-domain models may not transfer well to the clinical domain in all cases because of differences in language use. Domain-specific pretrained models can help, as can incorporating additional data of varying levels of specificity to the target task[16]—two techniques that we employed here. Our results attest to the utility of these two techniques, suggesting that there are sufficient commonalities in the way in which risk is expressed between these populations. However, future work may explore additional adaptation methods to further improve performance and generalizability across patient populations, such as recent methods for domain adaption of language models.[16] Furthermore, there is a selection bias toward higher-risk, more communicative individuals: Reddit users proactively reach out, while Caring Contacts participants can merely react (a core component of the intervention, designed to lower barriers to seeking help). This is reflected in part in the inverse class imbalance (5:1 vs 1:5), and raises the possibility of increased false positives (low precision) when applying the model to the lower-prevalence data. We addressed this issue via class weighting; additionally, the final fine-tuning step optimizes the decision threshold for the Caring Contacts dataset, counteracting the risk of excessive false positives. Our results confirm that including the Reddit data indeed results in no reduction in precision, as well as an overall performance improvement, compared to the setting where it is not included. However, this may still affect the generalizability of our results in unforeseen ways. Future work will include a prospective evaluation in any new setting or population to ensure acceptable levels of false positives before proceeding with production deployment.

In addition, it is important to note that we did not aim to assess patients' true level of suicidality, clinical needs, or likelihood to respond to clinical intervention—and therefore, the likelihood of improving outcomes (preventing suicide attempts or deaths). The expert-assigned data labels represent the degree to which the text expresses suicidality, but this may not match true risk: depending on their self-disclosure goals, individuals may over- or under-represent their distress levels[47,48]; in addition, SM and clinical text message-based settings differ in ways that affect the tradeoff between self-presentation and self-disclosure, such as anonymity, audience multiplicity (one-to-many vs one-to-one communication), and audience feedback (upvotes).[49] While suicide outcomes data could provide labels for true risk, the task of practical significance in our real-time triage setting is to make an initial determination of which messages warrant further investigation. Our models are not designed to determine clinical needs or recommend follow-up actions; rather, they are intended as a tool for clinicians to prioritize patients for detailed review and follow-up within the Caring Contacts intervention, using only what patients choose to share in written communication. In this way, our models empower clinicians to provide clinical care that is concordant with intervention guidelines (in terms of timeliness of follow-up). While the efficacy of Caring Contacts is supported by current evidence,[7] any actual reduction in attempts or deaths depends on other important factors, including effective intervention. Caring Contacts may not be appropriate or effective for all patient populations.

We have made several assumptions to estimate clinical utility. We assume that the CCVT dataset is a "gold standard" representing what human triage can achieve; however, response times were not a primary outcome of the trial. Staff used clinical judgment for response timing, but there were no explicit instructions for timing—except for messages indicative of an acute crisis, which were

addressed by immediately reaching out to participants via phone call. Other practical constraints were also at play, for example, work hours. After manual review, response times that seemed nonrepresentative were filtered out, but this does not fully alleviate the problem. If the original study protocol had included explicit timing instructions, there would likely be a more significant difference between response times for the two groups in this dataset. Second, in a real clinical practice setting, staff estimates of severity would likely be more fine-grained than "urgent" versus "not urgent". In this study, we grouped messages into just two urgency levels, resulting in the loss of some of this granularity and blurring the boundaries between groups. Third, the choice of summary statistic used to estimate workflow factors may influence utility calculations. For example, the median response times were lower (88.6 vs 116.2 min) compared to averages (106.2 vs 116.2 min). Here, the use of the median, rather than the average, did not substantially change calculated utilities; however, in practice, the choice of summary statistic may be meaningful and should be considered carefully. Finally, the trial was well-staffed to support the number of enrollees, and may represent a best-case scenario for staffing constraints which are likely to be more severe in settings where funding to support personnel is more limited. Nevertheless, we believe that our aim of *estimating* the clinical utility is achieved: even if it is only an estimate subject to several assumptions, it is more interpretable than other performance metrics that may mean very little to clinical stakeholders.

It is important to consider the implications of transfer learning-based, neural approaches for implementations in clinical production systems. In a phenomenon known as model drift, training data may become less representative of future observations, resulting in declining predictive performance over time. Updating datasets and retraining models periodically is effective in counteracting this problem. However, multiple data sources are involved with transfer learning, multiplying the effort required to keep models current. In addition, due to their complexity, deep learning models are computationally intensive—here, the BERT model had over 300 million trainable parameters, 250 000-fold more than the bag-of-words model at 1220. As message data grow, BERT models may require specialized hardware (Graphical or Tensor Processing Units) for training; although inference is fast compared to training, such hardware may also facilitate the rapid responsiveness appropriate for risk assessment applications. Any delay that is introduced may or may not have a practical impact on production systems, depending on the requirements of the system. In real-time clinical production systems, the improved predictive power of neural approaches must therefore be weighed carefully against computational resource limitations.

Future work

Further optimizations to the presented training process are possible. For example, it is possible that additional hyperparameter tuning, such as an exhaustive search for optimal hyperparameters in the Howard and Ruder learning rate selection strategy (target learning rates, division factor, number of iterations of learning rate growth, etc.), would further improve performance.

The best-performing model presented here would be expected to perform well in the clinical setting that produced the CCVT dataset. We are currently developing an informatics-supported pilot implementation of Caring Contacts, which includes this model as a clinical decision support component.

CONCLUSION

Advances in suicide prediction work using SM data are promising, but the growing divide between cutting-edge NLP research and the realities of clinical practice raises questions about the applicability of models emerging from this research to clinical settings. Yet, this is where such models have the highest potential to produce tangible improvements ethically and effectively. We demonstrated the feasibility of offsetting the limitations of the small size of clinical datasets with neural transfer learning using related, more readily available nonclinical data from a publicly available benchmark; further, we demonstrated practical value to clinical practice using a novel time utility metric. The use of automated language models for triage can thus enable the scaling of the Caring Contacts intervention to enroll more patients than traditional triage methods might allow. Thus, this work contributes toward bridging the historical implementation gap by translating state-of-the-art NLP advances to data from clinical settings and formally estimating utility toward improved risk assessment, paving the way for better clinical care and, ultimately, improved clinical outcomes.

FUNDING

This work was supported by the Garvey Institute for Brain Health Solutions Innovation Grant "Informatics-Supported Authorship for Caring Contacts (ISACC)". This work was in part supported by the Military Suicide Research Consortium, an effort supported by the Office of the Assistant Secretary of Defense for Health Affairs under Award No. W81XWH-10-2-0181. The views expressed herein are those of the author(s) and do not reflect the official policy or position of the U.S. Army Medical Department, Department of the Army, Department of Defense, the U.S. Government, or the Military Suicide Research Consortium.

AUTHOR CONTRIBUTIONS

HAB contributed to the idea for the work, conducted the experiments, and drafted and revised the manuscript. XD contributed to conducting the experiments and revised the article. AK contributed data and revised the article. KAC contributed to the idea for the work, contributed data, and revised the manuscript. TC contributed to the idea for the work and revised the article. All authors approved the article.

ACKNOWLEDGMENTS

The University of Maryland Reddit Suicidality Dataset was provided by Dr. Philip Resnik. We acknowledge the assistance of the American Association of Suicidology in making the dataset available.

CONFLICT OF INTEREST STATEMENT

None declared.

DATA AVAILABILITY

The University of Maryland Reddit Suicidality Dataset is available upon request to Dr. Philip Resnik from http://users.umiacs.umd.edu/~resnik/umd_reddit_suicidality_dataset.html.

The Caring Contacts Via Text dataset cannot be shared publicly for reasons of participant privacy and confidentiality. The data may be shared upon reasonable request to Dr. Katherine Anne Comtois.

REFERENCES

1. Stone DM, Simon TR, Fowler KA, *et al.* Trends in state suicide rates 1999-2016. *MMWR Morb Mortal Wkly Rep* 2018; 67 (22): 617–24.
2. Simon GE, Yarborough BJ, Rossom RC, *et al.* Self-reported suicidal ideation as a predictor of suicidal behavior among outpatients with diagnoses of psychotic disorders. *Psychiatr Serv* 2019; 70 (3): 176–83.
3. The Joint Commission. National Patient Safety Goal for suicide prevention. *R3 Rep* 2019: 1–6.
4. Stanley B, Brown GK, Brenner LA, *et al.* Comparison of the safety planning intervention with follow-up vs usual care of suicidal patients treated in the emergency department. *JAMA Psychiatry* 2018; 75 (9): 894–900.
5. Comtois KA, Kerbrat AH, DeCou CR, *et al.* Effect of augmenting standard care for military personnel with brief caring text messages for suicide prevention. *JAMA Psychiatry* 2019; 76 (5): 474–83.
6. Reger MA, Luxton DD, Tucker RP, *et al.* Implementation methods for the caring contacts suicide prevention intervention. *Prof Psychol Res Pract* 2017; 48 (5): 369–77.
7. Skopp NA, Smolenski DJ, Bush NE, *et al.* Caring contacts for suicide prevention: a systematic review and meta-analysis. *Psychol Serv* 2022; 20 (1): 74–83.
8. Burkhardt HA, Laine M, Kerbrat A, *et al.* Identifying opportunities for informatics-supported suicide prevention: the case of Caring Contacts. In: AMIA annu symp proc; 2022.
9. Halevy A, Norvig P, Pereira F. The unreasonable effectiveness of data. *IEEE Intell Syst* 2009; 24 (2): 8–12.
10. Gururangan S, Marasović A, Swayamdipta S, *et al.* Don't stop pretraining: adapt language models to domains and tasks. In: proceedings of the 58th annual meeting of the Association for Computational Linguistics; 2020: 8342–60. doi:10.18653/v1/2020.acl-main.740
11. Payne TH, Corley S, Cullen TA, *et al.* Report of the AMIA EHR-2020 Task Force on the status and future direction of EHRs. *J Am Med Inform Assoc* 2015; 22 (5): 1102–10.
12. Adler-Milstein J, Jha AK. Health information exchange among U.S. hospitals: Who's in, who's out, and why? *Healthc (Amst)* 2014; 2 (1): 26–32.
13. Coppersmith G, Leary R, Crutchley P, *et al.* Natural language processing of social media as screening for suicide risk. *Biomed Inform Insights* 2018; 10: 1–11. doi:10.1177/1178222618792860
14. Shing H-C, Nair S, Zirikly A, *et al.* Expert, crowdsourced, and machine assessment of suicide risk via online postings. In: *proceedings of the fifth workshop on computational linguistics and clinical psychology: from keyboard to clinic*. Stroudsburg, PA: Association for Computational Linguistics; 2018: 25–36. doi:10.18653/v1/W18-0603
15. Zirikly A, Resnik P, Uzuner Ö, *et al.* CLPsych 2019 shared task: predicting the degree of suicide risk in reddit posts. In: proc sixth work comput linguist clin psychol; 2019: 24–33.
16. Laparra E, Mascio A, Velupillai S, *et al.* A review of recent work in transfer learning and domain adaptation for natural language processing of electronic health records. *Yearb Med Inform* 2021; 30 (1): 239–44.
17. Weiss K, Khoshgoftaar TM, Wang DD. *A Survey of Transfer Learning*. London, United Kingdom: Springer International Publishing; 2016. doi:10.1186/s40537-016-0043-6.
18. Naseem U, Lee BC, Khushi M, *et al.* Benchmarking for public health surveillance tasks on social media with a domain-specific pretrained language model. In: proceedings of NLP power! The first workshop on efficient benchmarking in NLP; 2022. http://arxiv.org/abs/2204.04521
19. Ji S, Zhang T, Ansari L, *et al.* MentalBERT: publicly available pretrained language models for mental healthcare. In: proceedings of the thirteenth language resources and evaluation conference; 2021. https://github.com/Inusette/
20. Lee J, Yoon W, Kim S, *et al.* BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36 (4): 1234–1240.
21. Howard J, Ruder S. Universal language model fine-tuning for text classification. In: ACL 2018—56th annu meet Assoc Comput Linguist proc conf (long pap), Volume 1; 2018: 328–39. doi:10.18653/v1/p18-1031
22. Seneviratne MG, Shah NH, Chu L. Bridging the implementation gap of machine learning in healthcare. *BMJ Innov* 2020; 6 (2): 45–7.
23. Plana D, Shung DL, Grimshaw AA, *et al.* Randomized Clinical Trials of Machine Learning Interventions in Health Care. *JAMA Netw Open* 2022; 5 (9): e2233946.
24. Hernandez-Boussard T, Lundgren MP, Shah N. Conflicting information from the Food and Drug Administration: missed opportunity to lead standards for safe and effective medical artificial intelligence solutions. *J Am Med Inform Assoc* 2021; 28 (6): 1353–5.
25. Jung K, Kashyap S, Avati A, *et al.* A framework for making predictive models useful in practice. *J Am Med Informatics Assoc* 2021; 28 (6): 1149–58.
26. Bayati M, Braverman M, Gillam M, *et al.* Data-driven decisions for reducing readmissions for heart failure: general methodology and case study. *PLoS One* 2014; 9 (10): e109264.
27. Barthel M, Stocking G, Holcomb J, *et al.* Nearly eight-in-ten reddit users get news on the site; 2016. www.pewresearch.org.
28. Hu Y, Talamadupula K, Kambhampati S. Dude, srsly?: The surprisingly formal nature of Twitter's language. *ICWSM* 2021; 7 (1): 244–53.
29. De Choudhury M, Sharma SS, Logar T, *et al.* Gender and cross-cultural differences in social media disclosures of mental illness. In: proc ACM conf comput support coop work CSCW 2017: 353–369. doi:10.1145/2998181.2998220
30. Devlin J, Chang M-W, Lee K, *et al.* BERT: pre-training of deep bidirectional transformers for language understanding. In: proceedings of the 2019 conference of the North American chapter of the Association for Computational Linguistics: human language technologies (long and short papers), Volume 1. Association for Computational Linguistics; 2019: 4171–86. doi:10.18653/v1/N19-1423
31. Bird S, Loper E, Klein E. *Natural Language Processing with Python*. Sebastopol, CA: O'Reilly Media Inc.; 2009.
32. Cawley GC, Talbot NLC. On over-fitting in model selection and subsequent selection bias in performance evaluation. *J Mach Learn Res* 2010; 11: 2079–107.
33. Pedregosa F, Varoquaux G, Gramfort A, *et al.* Scikit-learn: machine learning in Python. *J Mach Learn Res* 2011; 12: 2825–30.
34. Wolf T, Debut L, Sanh V, *et al.* HuggingFace's transformers: state-of-the-art natural language processing. In: proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations; 2019: 38–45. doi:10.18653/v1/2020.emnlp-demos.6.
35. Coppersmith G, Dredze M, Harman C. Quantifying mental health signals in Twitter. In: *proceedings of the workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality*. Stroudsburg, PA: Association for Computational Linguistics; 2014: 51–60. doi:10.3115/v1/W14-3207.
36. Shen JH, Rudzicz F. Detecting anxiety through Reddit. In: proceedings of the fourth workshop on computational linguistics and clinical psychology—from linguistic signal to clinical reality. Stroudsburg, PA: Association for Computational Linguistics; 2017: 58–65. doi:10.18653/v1/W17-3107
37. De Choudhury M, Counts S, Horvitz EJ, *et al.* Characterizing and predicting postpartum depression from shared Facebook data. In: proc ACM conf comput support coop work CSCW; 2014: 626–38. doi:10.1145/2531602.2531675
38. Resnik P, Garron A, Resnik R. Using topic modeling to improve prediction of neuroticism and depression in college students. In: EMNLP 2013—2013 conf empir methods nat lang process proc conf; 2013: 1348–1353.
39. Gomes de Andrade NN, Pawson D, Muriello D, *et al.* Ethics and artificial intelligence: suicide prevention on Facebook. *Philos Technol* 2018; 31 (4): 669–84.
40. Lee N. Trouble on the radar. *Lancet* 2014; 384 (9958): 1917.

41. Hull TD, Malgaroli M, Connolly PS, *et al.* Two-way messaging therapy for depression and anxiety: longitudinal response trajectories. *BMC Psychiatry* 2020; 20 (1): 297.
42. Burkhardt H, Pullmann M, Hull T, *et al.* Comparing emotion feature extraction approaches for predicting depression and anxiety. In: proceedings of the eighth workshop on computational linguistics and clinical psychology. Stroudsburg, PA: Association for Computational Linguistics; 2022: 105–15. doi:10.18653/v1/2022.clpsych-1.9.
43. Barnett I, Torous J. Ethics, transparency, and public health at the intersection of innovation and Facebook's suicide prevention efforts. *Ann Intern Med* 2019; 170 (8): 565–6. doi:10.7326/M19-0366
44. Chancellor S, Birnbaum ML, Caine ED, *et al.* A taxonomy of ethical tensions in inferring mental health states from social media. In: FAT* 2019—proc 2019 conf fairness, accountability, transpar; 2019: 79–88. doi:10.1145/3287560.3287587.
45. Horvitz E, Mulligan D. Data, privacy, and the greater good. *Science* 2015; 349 (6245): 253–5.
46. Singer N. In screening for suicide risk, facebook takes on tricky public health role. *New York Times*. 2018. https://nyti.ms/2RkXJCn. Accessed September 29, 2020.
47. Bazarova NN, Choi YH. Self-disclosure in social media: extending the functional approach to disclosure motivations and characteristics on social network sites. *J Commun* 2014; 64 (4): 635–57. doi:10.1111/jcom.12106
48. Bazarova NN, Taft JG, Choi YH, *et al.* Managing impressions and relationships on Facebook: self-presentational and relational concerns revealed through the analysis of language style. *J Lang Soc Psychol* 2013; 32 (2): 121–41.
49. Schlosser AE. Self-disclosure versus self-presentation on social media. *Curr Opin Psychol* 2020; 31: 1–6.