Psychological Medicine

cambridge.org/psm

Detecting suicide risk among U.S. servicemembers and veterans: a deep learning approach using social media data

Kelly L. Zuromski[1,2], Daniel M. Low[3,4,*], Noah C. Jones[1,5,*], Richard Kuzma[1], Daniel Kessler[1], Liutong Zhou[6], Erik K. Kastman[1,7], Jonathan Epstein[7], Carlos Madden[7], Satrajit S. Ghosh[3,4], David Gowel[7] and Matthew K. Nock[1]

[1]Department of Psychology, Harvard University, Cambridge, MA, USA; [2]Franciscan Children's, Brighton, MA, USA; [3]Speech and Hearing Bioscience and Technology Program, Harvard Medical School, Boston, MA, USA; [4]McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge MA; [5]MIT Media Lab, Massachusetts Institute of Technology, Cambridge, MA, USA; [6]Machine Learning Solutions Lab, Amazon Web Services, New York, NY, USA and [7]RallyPoint Networks, Inc., Boston, MA, USA

Original Article

*Equal contribution.

Cite this article: Zuromski KL *et al* (2024). Detecting suicide risk among U.S. servicemembers and veterans: a deep learning approach using social media data. *Psychological Medicine* **54**, 3379–3388. https://doi.org/10.1017/S0033291724001557

Received: 22 December 2023
Revised: 22 May 2024
Accepted: 30 May 2024
First published online: 9 September 2024

Keywords:
machine learning; military; natural language processing; suicide; veterans

Corresponding author:
Kelly L. Zuromski;
Email: kelly_zuromski@fas.harvard.edu

Abstract

Background. Military Servicemembers and Veterans are at elevated risk for suicide, but rarely self-identify to their leaders or clinicians regarding their experience of suicidal thoughts. We developed an algorithm to identify posts containing suicide-related content on a military-specific social media platform.

Methods. Publicly-shared social media posts ($n = 8449$) from a military-specific social media platform were reviewed and labeled by our team for the presence/absence of suicidal thoughts and behaviors and used to train several machine learning models to identify such posts.

Results. The best performing model was a deep learning (RoBERTa) model that incorporated post text and metadata and detected the presence of suicidal posts with relatively high sensitivity (0.85), specificity (0.96), precision (0.64), F1 score (0.73), and an area under the precision-recall curve of 0.84. Compared to non-suicidal posts, suicidal posts were more likely to contain explicit mentions of suicide, descriptions of risk factors (e.g. depression, PTSD) and help-seeking, and first-person singular pronouns.

Conclusions. Our results demonstrate the feasibility and potential promise of using social media posts to identify at-risk Servicemembers and Veterans. Future work will use this approach to deliver targeted interventions to social media users at risk for suicide.

Suicide is a leading cause of death among U.S. military Servicemembers and Veterans. Since 9/11, almost four times as many Servicemembers have died by suicide than in combat (Suitt, 2021) and the Veteran suicide rate has exceeded that of the U.S. general population (Kang et al., 2015). Unfortunately, many military personnel who may be at risk for suicide (e.g. those with diagnosable mental disorders) do not receive mental health treatment (Colpe et al., 2015; Hoge et al., 2004). In a study of Army soldiers, only about 25% of soldiers who died by suicide were seen by a mental health professional in the month before their death, with suicide risk undocumented (and likely unknown) for approximately 85% of these individuals (Ribeiro et al., 2017). There are many reasons suicidal military personnel may not seek treatment, including stigma of mental health care, or structural barriers such as availability of treatment (Zuromski et al., 2019). Novel approaches are critically needed to ensure military personnel receive the help they need, particularly outside of traditional health care settings.

One way to improve resource availability for at risk individuals is to use social media. Self-disclosures related to mental health and suicide are relatively common on social media (e.g. De Choudhury & De, 2014; Low et al., 2021; Naslund, Bondre, Torous, & Aschbrenner, 2020), which may provide an opportunity to identify people who would likely benefit from help. Prior studies have used a range of machine learning methods to identify suicide risk in social media posts (for review see Castillo-Sánchez et al., 2020; Homan et al., 2022). This work has demonstrated the feasibility of using social media data for detecting users' suicide risk on Twitter (e.g. MacAvaney, Mittu, Coppersmith, Leintz, & Resnik, 2021; O'Dea et al., 2015), Reddit (e.g. De Choudhury & De, 2014; Jones, Jaques, Pataranutaporn, Ghandeharioun, & Picard, 2019), Facebook (e.g. Ophir, Tikochinski, Asterhan, Sisso, & Reichart, 2020), and Instagram (e.g. Lekkas, Klein, & Jacobson, 2021).

Little research has specifically examined Servicemembers' and Veterans' mental health disclosures on social media. One study examined the content of social media posts of suicide decedents in the U.S. military to identify risk factors that may have suggested the individual was at risk (Bryan et al., 2018). However, to our knowledge, no studies have prospectively

identified military personnel on social media who may be at elevated suicide risk. One reason for the limited research in this area may be that military personnel are not likely to help-seek or disclose mental health problems on popular social media sites like Facebook (Teo et al., 2018). Given the unique stressors and challenges that Servicemembers and Veterans face, many individuals may be more comfortable sharing mental health concerns with similarly situated peers who understand the military's unique culture.

In the current study, we developed an algorithm that flags at-risk users on a military social media platform called RallyPoint. This social media site was developed in response to the lack of dedicated social and professional digital community for military personnel. RallyPoint has nearly 2 million active users and provides a platform where Servicemembers and Veterans can connect and get support around topics such as joining the military, deployment, transition to civilian life, employment, and mental health. In this study, we collected and coded a large corpus of RallyPoint posts to develop a machine learning model to detect suicide-related content risk and to examine how users share about their suicide risk on the site.

Method

Data collection

Public posts (~5.3 million) shared on RallyPoint between September 2013 – March 2020 were used in this study. Posts were pulled from RallyPoint's Redshift database hosted on Amazon Web Services. This study was approved by Harvard University's Institutional Review Board (IRB19-1260) and by the Department of Defense's Human Research Protection Office (FWR20200173X). The authors assert that all procedures contributing to this work comply with the ethical standards of the relevant national and institutional committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2008.

Human data labeling

Because suicidal thoughts and behaviors (STB) are low base rate behaviors and come up in a minority of posts on RallyPoint, we first identified posts most likely to contain STB content to prioritize for human labeling. To do so, we used the Snorkel weak supervision library (Ratner et al., 2020). Weak supervision employs imperfect labeling functions, which are heuristic rules or noisy classifiers to estimate the likelihood of posts containing STB language. We identified a dictionary of keywords related to suicide (e.g. 'die', 'pills') and suicide risk factors (e.g. 'ptsd', 'hopeless') and labeling functions were designed to flag posts containing these keywords. In addition, we designed labeling functions that identified patterns indicative of non-STB posts. For instance, posts containing advertisement links or advertising phrases were automatically classified as non-STB. Next, these labeling functions were applied to the entire unlabeled dataset. Snorkel then used a generative model to combine the outputs of these functions, estimating a probabilistic label for each post. Posts with high likelihood of STB content were prioritized for review by our labeling team.

For human data labeling, we developed a codebook that contained a list of STB-related words (e.g. thoughts of death, passive or active suicidal ideation, plan, attempt) and examples. Given the low base rate of STB, we coded posts as present/absent for any STB content, rather than focusing on any one type of STB (e.g. active suicidal ideation). We focused exclusively on explicit mentions of the user's personal experiences with STB, not posts containing only descriptions of known risk factors for suicide (e.g. depression, PTSD). Next, we coded a subset of posts ($n = 500$), which allowed us to make revisions and additions to our codebook, including adding a list of colloquial and military-specific phrases and terms related to STB. For example, we included the phrase 'I'm afraid I'll become one of the 22 a day,' as an example of STB language, which refers to the commonly referenced statistic in the military community that 22 Veterans die by suicide every day. We also included examples of firearm-related jargon (e.g. 'I thought about eating a round,' 'I had the 0.45 at my temple') and noted examples of ambiguous STB-related phrases, in which the overall context of the post is needed to determine whether it is a STB post or not (e.g. 'I want to sleep forever'). We also added coding rules for specific cases that were frequently observed (e.g. descriptions of 'suicide bombings' were coded non-STB; descriptions of a friend's or fellow soldier's suicide were coded non-STB unless the post also contained description of the user's STB).

After finalizing the codebook, 2000 additional posts were double-coded by our team for the overall presence or absence of STB content using Amazon Web Services' SageMaker labeling service. Additional labeling ($n = 6000$) was conducted by third-party annotators, who received training on our STB codebook and supervision from our team to ensure labeling consistency. Inter-rater reliability was calculated using the package *IRR* in R (Gamer, Lemon, Fellows, & Singh, 2019). Substantial agreement was achieved for both groups (Krippendorff's alpha = 0.73 and 0.65, respectively; Hughes, 2021). All coding discrepancies, from both internal and third-party coded posts, were resolved using consensus coding involving at least one Ph.D.-level member of the team. After removal of duplicate posts, a total of 7967 posts were labeled.

Data preprocessing

Data augmentation

Given the high imbalance of STB to non-STB posts, we augmented the text data after labeling was complete. Specifically, we hand-crafted additional STB posts ($n = 230$) using three methods. First, we generated new STB posts that were similar in style and content to the labeled STB posts. Second, we referenced STB-related phrases from other social media work conducted by our group and generated posts similar in content to these external sources. Lastly, we also generated additional non-STB posts ($n = 252$) to reduce the risk of introducing bias into the model.

We split the final augmented dataset ($N = 8449$ posts) into a training, validation, and test set, seeking a 70%/10%/20% split across the datasets with a consistent (stratified) ratio of non-STB to STB classes. All posts by a unique user were only included in one of the sets to avoid data leakage. We removed any augmented or duplicated posts from the final test set so it reflected real-world posts only. This resulted in a 76.7% / 8.2% / 15.1% breakdown between the training, validation, and test sets with an approximately 10:1 ratio of non-STB posts to STB posts. The full breakdown is found in online Supplemental Table 1.

Processing post text and metadata

We extracted post and user metadata features for use in analyses: post type (status update, comment, or question), tags (user-

https://doi.org/10.1017/S0033291724001557 Published online by Cambridge University Press

defined labels selected from a list on RallyPoint, e.g. humor, technology, ptsd), reputation (a RallyPoint site metric that measures how engaged and influential a user is, both in terms of how active a user is on the site and how much attention their posts receive from other users), and contact size (the number of users each user has in their RallyPoint contact list). In our machine learning models that included metadata features, tags and post type (text variables) were concatenated to title and body strings. We cleaned text data by removing hyperlinks within posts and any hypertext markup language (html). For the continuous numerical metadata features (reputation and contact size) we normalized values between 0 to 1 scale.

Classic machine learning models cannot operate directly on raw text and require that text be transformed into vectors. We used the Python scikit-learn (Pedregosa et al., 2011) implementation of term frequency-inverse document frequency (TF-IDF) (Sparck Jones, 1972) to vectorize text from posts and the title of the post (if available). TF-IDF gives statistical importance to each word within a document with respect to the entire corpus of documents. If a word is rare within the entire corpus of documents but prevalent within a particular document, then that word will likely have a high TF-IDF score for that document relative to others in the corpus.

Model development

Classic machine learning models

Initially, we tested both linear (logistic regression; LogReg) and nonlinear (light gradient boosting machines; LGBM) models. For both LogReg and LGBM, we tested two types of models: those including only the extracted text features (using TF-IDF, as described in *Processing post text and metadata*) from users' posts (Text) and those that also included user metadata features (Text + Metadata). We performed hyperparameter tuning using Gridsearch on a 5-fold cross-validation on the combined training and validation sets on the following hyperparameters: for TF-IDF text features, maximum feature size (256, 2048, None); for logistic regression, regularization type (L1, L2), regularization strength (C: 0.1, 0.3, 0.6, 1); for LGBM, max. depth (10, 20, None), min._data_in_leaf (10, 20, 40), min_child_weight (0.01, 0.001, 0.0001) and feature fraction (0.1, 0.5, 1). To deal with the class imbalance problem, we tested both random oversampling and class weights to all models to penalize prediction errors on the minority class more heavily.

A limitation of using TF-IDF to extract text features for these models is that it does not account for contextualized semantic context. The TF-IDF features of two sentences such as 'this test is too hard, I give up' and 'this life is too hard, I give up' turn out to be extremely similar in the feature vector space because they share similar words, even though the semantic meaning of them is quite different. This problem is addressed in deep learning models.

Deep learning models

Compared with classical approaches, a deep learning approach using Bidirectional Encoder Representations from Transformers (BERT) family of models has huge advantages in capturing lexical morphology, syntax, and contextualized semantics at the same time. We used one such model, RoBERTa, which tends to give good performance over other commonly used models such as BERT (Liu et al., 2019). Deep learning models such as RoBERTa are first trained on large corpora (e.g. the entire Wikipedia, a large book corpus) in a self-supervised manner (i.e. predicting masked words within the input instead of a ground-truth label) (Liu et al., 2019). This process allows the model to train over massive unlabeled datasets and create representations that can capture semantic, morphological, and syntactic information of a given document which can then be used on many downstream tasks (e.g. question-answering, text classification).

To develop a text-only RoBERTa model (Text), we finetuned the RoBERTa-base pretrained model through the Python *transformers* package (Wolf et al., 2020) on the training set. Finetuning familiarizes the model with the style of language used in RallyPoint posts. We used Optuna v3.3.0 (Akiba, Sano, Yanase, Ohta, & Koyama, 2019) for hyperparameter search on the labeled training set of STB and non-STB posts to perform the binary classification task that is the focus of our study. Optuna adaptively selects best parameter combinations (i.e. trials) by focusing on values where hyperparameters are giving the best results and implements early stopping to prune unpromising trials for high-performing optimization, which is much faster than an exhaustive grid search. We tested another RoBERTa model that included metadata (Text + Metadata model), using the multimodal toolkit (Gu & Budhkar, 2021).

On both the Text and Text + Metadata deep learning models, we searched among the following hyperparameters based on the default options set by the transformers package, removing batch size to reduce GPU memory usage, and adding weight decay: learning rate (1e-6 – 1e-4), and epoch size (1–4), and weight decay (1e-10 – 1e-3) for 10 trials evaluated on the validation set. RoBERTa was chosen because it showed the highest performance on the GLUE leaderboard (performance on nine tasks) among the models available in the multimodal toolkit (e.g. BERT, XLM, DistilBERT). We also tested random oversampling and class weights to treat the class imbalance problem for the RoBERTa models.

Model evaluation

Using the test set, we computed precision (i.e. positive predictive value), sensitivity (i.e. recall), specificity, and F1 score to evaluate our final model performance. The F1 score is the harmonic mean of precision and sensitivity that is only high when both precision and sensitivity are high; however, it uses a threshold of 0.5 which may be suboptimal. We do not use accuracy as a model evaluation metric because it is biased by class imbalance; that is, a model could achieve close to 90% accuracy merely by predicting 'non-STB' for every post. Similarly, interpretation of the area under the ROC curve (AUC) is biased by class imbalance because AUC is inflated by predicting most of the negatives (non-STB posts) correctly and there are many negatives in this dataset, which is not as important as predicting the positives (STB posts) correctly. Therefore, we also provide a precision-recall curve, which provides a more realistic prediction of future classification performance because it evaluates the fraction of true positives among positive predictions at different thresholds (Saito & Rehmsmeier, 2015).

Semantic comparison of STB v. non-STB posts

Deep learning models are inherently limited in their interpretability. As such, we conducted additional analyses to identify words that are more likely to occur in the STB posts than in the

https://doi.org/10.1017/S0033291724001557 Published online by Cambridge University Press

non-STB posts using posts from the training set, which can help explain the statistical regularities the successful models may be using to classify the two groups of posts. These analyses also help us to understand how Servicemembers and Veterans are talking about their STB on the site. To compare STB v. non-STB posts, we used the *scattertext* package (Kessler, 2017) on selected post text and metadata (i.e. concatenation of title, body, tags, post type). We used the scaled f-score, which provided similar results to the more standard weighted log odds ratio (Monroe, Colaresi, & Quinn, 2008), but with better visibility of the difference between words. We also conducted descriptive analyses to examine differences in metadata (e.g. post type, tags) between STB and non-STB posts.

Results

Characterizing the sample

Our final sample included 8449 posts (7967 posts from RallyPoint, along with an additional 482 generated by our team using the data augmentation methods described in the Method). The 7697 RallyPoint posts were from 2747 unique users who were mostly male (86.0%) and 55 or older (28.7%). The posts were predominantly from Veteran users (68.9%; Servicemembers 27.5%) who were/are in the Army (66.3%). Users making STB posts were more often Veterans (79.8%) than users making non-STB posts (67.5%), but otherwise demographics between users making STB and non-STB posts were similar. Compared to the overall RallyPoint user base, which is split about equally between Veterans and Servicemembers, our sample included more Veterans and also tended to be older than the average RallyPoint user. See online Supplemental Table 2 for detailed demographic and profile information on users in our sample and of RallyPoint in general.

Evaluation of final model performance

We tested and compared several models (Table 1) and our best performing model was a RoBERTa-based model incorporating posts' text and metadata. The confusion matrix for the final model is shown in Table 2. This model achieved a sensitivity of 0.85 (i.e. 85% of all STB posts were correctly flagged), a specificity of 0.96 (i.e. 96% of non-STB posts were correctly classified as non-STB), a precision of 0.64 (i.e. of the posts predicted to be STB, 64% actually were), and an F1 score of 0.73 (harmonic mean of sensitivity and precision). The ROC and precision-recall curve plots for the final model are shown in Fig. 1.

Error analysis

We manually examined posts that were incorrectly classified by the final model. We selected all 16 false negative posts to inspect (i.e. test set posts that our human labelers identified as STB but were classified by the model as non-STB). Most of these posts were ambiguous and not straightforward STB posts. For example, in one post, the user described that they would never tell their doctor that they have had suicidal thoughts. This is an unclear post; it's possible that the user has had suicidal thoughts and is describing that they would not disclose these. Alternatively, this post could be hypothetical and describing a possible future event. Thus, even though this post was coded by our team as an STB post, it was not a clear-cut case of STB language and it is not surprising that the model missed similarly ambiguous cases.

Of the false positive posts (i.e. posts that our coding team identified as non-STB but the model classified as STB), we selected a random subset of 16 posts to further inspect. These posts all contained descriptions of risk factors for suicide (e.g. PTSD, depression, substance use) or suicide-related words (e.g. discussions of being bullied and encouraged to kill themselves). However, none of the false positive posts that were reviewed contained explicit descriptions of STB as defined in our codebook.

Semantic analysis of STB v. non-STB posts

We conducted additional analyses to understand which words would help a model to distinguish between STB and non-STB posts using posts from the training set. Results are shown in Fig. 2. This figure shows words that are more likely to appear in STB posts than non-STB posts, which include words related to suicide (e.g. 'suicide', 'attempted', 'thoughts', 'ideations', 'kill', 'pills'), suicide risk factors (e.g. 'hopeless', 'burden', 'depression', 'cry', 'ptsd'), help-seeking (e.g. 'help', 'talk', 'conversation', 'asist' [Applied Suicide Intervention Skills Training]), first-person singular pronouns ('I', 'me', 'myself', 'm' [contraction for 'am']) and negations ('not' and the contracted form 't', 'don't', 'can't', 'didn't', 'nobody'). Words that were more likely to appear in non-STB posts were related to third-person singular pronouns ('he', 'him', 'she', 'his'), war and military history ('war', 'soldiers', 'died', 'killed', 'Japanese', 'attack'), and religion ('sin', 'john', 'bible'). Also more likely to occur in non-STB posts is the use of asterisks, which are often used to add emphasis in posts on RallyPoint or used to censor profanity. Examples posts containing words frequently found in STB posts are provided in Table 3.

We also examined words that characterize RallyPoint posts, in general, compared to a general English corpus, which are listed in

Table 1. *Model performance*

Model	Sensitivity	Specificity	Precision	F1	ROC AUC	PR AUC
LogReg text	0.44	0.92	0.35	0.39	0.68	0.29
LogReg text + metadata	0.53	0.91	0.36	0.43	0.72	0.34
LGBM text	0.54	0.93	0.42	0.47	0.73	0.38
LGBM text + metadata	0.46	0.96	0.53	0.49	0.71	0.50
RoBERTa text	0.73	0.96	0.65	0.69	0.85	0.69
*RoBERTa text + metadata	0.85	0.96	0.64	0.73	0.98	0.84

Note. LogReg, Logistic Regression; LGBM, Light Gradient Boosting Machine; RoBERTa, Robustly Optimized BERT Pretraining Approach. Text: text features from the title and body of RallyPoint posts were included in the model; Text + Metadata: text features and metadata features were included. ROC AUC: Area under the receiver operating curve; PR AUC: Area under the precision-recall curve. *Final, best-performing model.

https://doi.org/10.1017/S0033291724001557 Published online by Cambridge University Press

Table 2. Confusion matrix for the final, best performing RoBERTa model

		Predicted post classification	
		Non-STB	STB
Actual post classification	Non-STB	True negatives: 87% ($N = 1117$; 96% of actual Non-STB posts)	False positives: 4% ($N = 52$; 4% of actual Non-STB posts)
	STB	False negatives: 1% ($N = 16$; 15% of actual STB posts)	True positives: 7% ($N = 94$; 85% of actual STB posts)

Note. STB, human labelers identified words or phrases relating to STB in the post; Non-STB, no STB content was identified by human labelers in the post.

Fig. 2. Identifying characteristic words in RallyPoint posts can help us understand why models trained in the current dataset may not generalize to datasets without similar word patterns. Compared to a general English corpus, RallyPoint posts are more likely to contain words related to the military ('veterans', 'ncos' [non-commissioned officers], 'vets'), war and politics ('iwo jima', 'obama'), and physical/mental health ('ptsd' [Posttraumatic Stress Disorder], 'suicidal', 'tbi' [traumatic brain injury]).

See an interactive version of Fig. 2 to view scores and additional words (https://tinyurl.com/suicidal-vs-non-suicidal).

Descriptive analysis of metadata features in STB v. non-STB posts

In addition to comparing text content, we conducted descriptive analyses to identify differences in metadata features between STB v. non-STB posts. Specifically, we examined frequency of post type (i.e. status update, comment, or question), user-generated tags, contact size, and reputation (a RallyPoint site metric of how engaged and influential a user is). For both STB and non-STB posts, the most common post type was comment, meaning that the user's post was in response to another user's status update. The most common user-generated tag on STB posts was 'suicide,' followed by 'PTSD,' which were used less frequently in non-STB posts. Full results for post type and tags can be found in online Supplemental Table 3. Users who made STB compared to non-STB posts had a similar number of RallyPoint contacts, but users who made non-STB posts tended to have higher reputation scores (online Supplemental Figure 1).

Discussion

There are two main findings in the current study. First, our results demonstrate the feasibility of developing and validating a well-performing algorithm that can detect suicide-related content using social media data. Second, we identified key differences in STB compared to non-STB posts, which sheds light on how military Servicemembers and Veterans share about their suicide risk on social media. We expand on each finding below.

We developed and validated a military-specific risk algorithm that accurately detects suicidal content in posts on a social media platform. Although similar models have been developed on other social media platforms (e.g. Castillo-Sánchez et al., 2020), none have been specific to military personnel, who are at particularly high suicide risk and experience significant barriers to help-seeking. Thus, our study fills a gap in the literature and demonstrates the utility of harnessing social media to identify at-risk

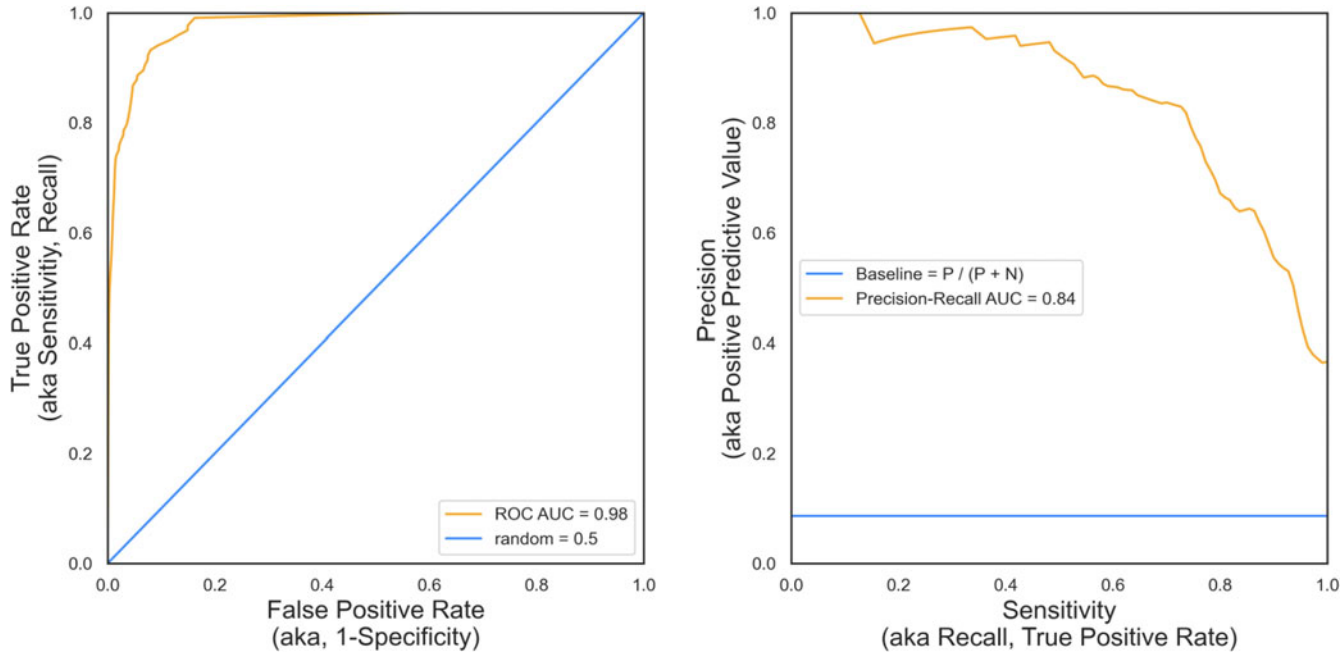

Figure 1. ROC (left) and precision-recall (right) plots for the final, best performing RoBERTa model classifying STB v. non-STB posts. Because ROC is biased by class imbalance, precision-recall is a more informative metric for detecting the minority group (i.e. STB posts) under class imbalance. Baseline in the precision-recall curve is the proportion of positive (STB posts) examples in our data which would be the result for a random baseline.

https://doi.org/10.1017/S0033291724001557 Published online by Cambridge University Press

https://doi.org/10.1017/S0033291724001557 Published online by Cambridge University Press

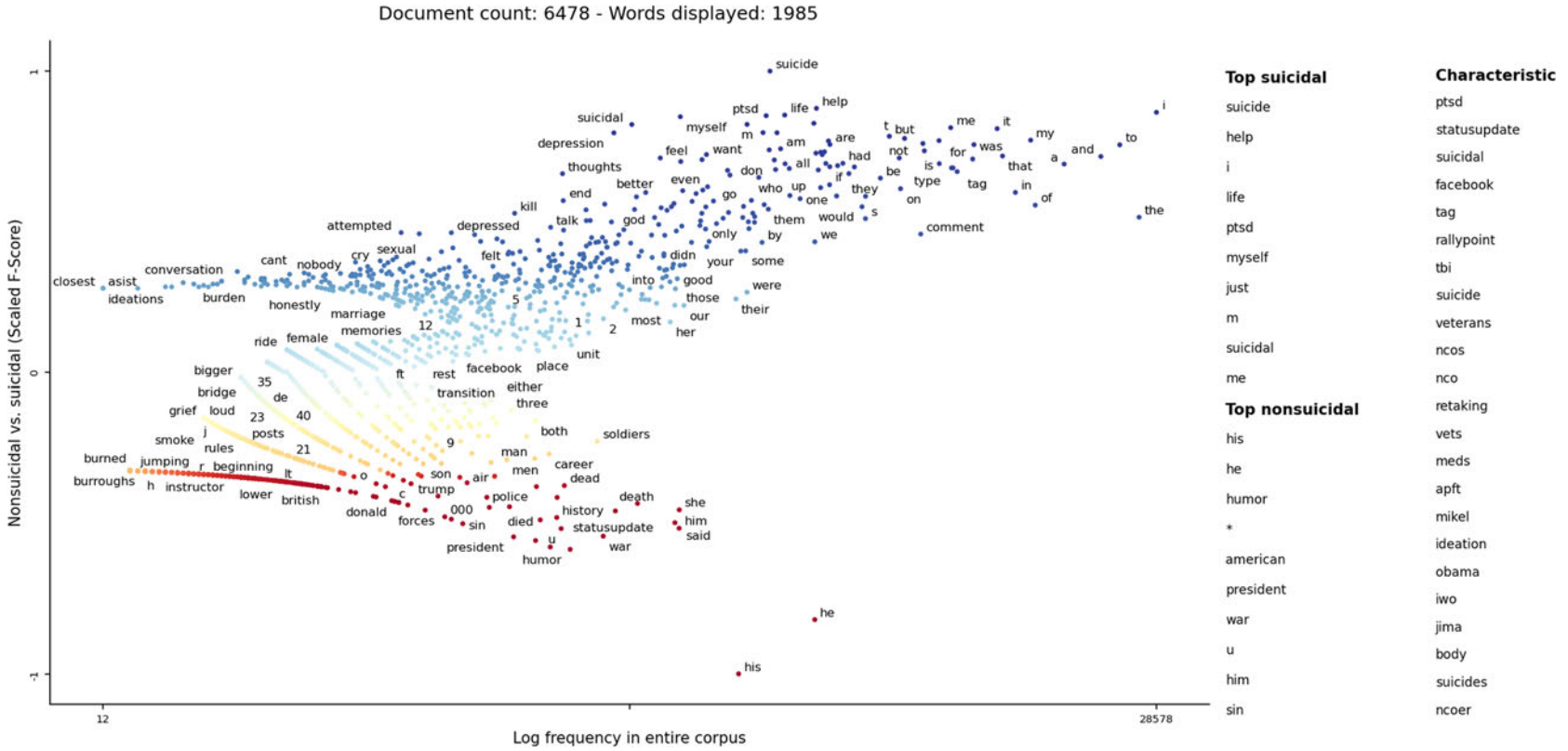

Figure 2. Semantic analysis of STB v. non-STB posts. This figure shows words that are more common in STB ('Top suicidal') and non-STB ('Top nonsuicidal') posts from the training set ($N = 6478$). Each dot represents a word. The x-axis represents word frequencies in the entire corpus of RallyPoint posts, whereas the y-axis represents how common a given word is based on post type, with 1 representing words very common in STB posts and −1 representing words very common in non-STB posts. The words under 'Characteristic' represent words that characterize the entire corpus of RallyPoint posts (including both STB and non-STB posts) in comparison to a general English Corpus. Only unique words that appear at least 7 times in the corpus are shown in the figure ($n = 1985$). Note: 'm' in the 'Top Suicidal' list represents the contraction for 'am'; * in the 'Top nonsuicidal' list reflects use of asterisks to add emphasis or censor profanity in posts. See an interactive version of Fig. 2 to view scores and additional words: https://tinyurl.com/suicidal-vs-non-suicidal

Table 3. Excerpts from posts containing words occurring frequently in STB posts

Common words	Excerpts from STB posts
Suicide	I have contemplated **suicide** twice in the past year and still do. I have been hospitalized twice once for suicidal idealization and once for attempted **suicide.** I've had these **suicide** thoughts for about 6 months now and it's getting worse
Risk factors (e.g. depression, PTSD)	I have had a troubled life, aside from the abuse the **sexual abuse** developed into **depression**, **suicidal tendencies,** and **PTSD**. Can't deal with the **shame**, **guilt**, and **depression** (possible **manic**) I now have very bad **PTSD** and really want to end it all. Someone please help
Help-seeking (e.g. help, therapy)	FINALLY found a decent therapist/LCSW to **help** me through the bad times **Therapy, meds**, and a loving husband that has held me when I'm at my worst has **helped** me to recover.
Just	Sometimes I **just** want to go jump off a bridge And I **just** feel I have no purpose
Negations	I **don't** think anybody would miss me, even if I **didn't** die in active duty and it was just my own doing. [It's] easier to keep it private. Many had **never** seen me cry

Note. STB, suicidal thoughts and behaviors.

military personnel. Given most Servicemembers and Veterans who die by suicide do not seek out mental health care in the months preceding their deaths (Colpe et al., 2015; Ribeiro et al., 2017), creative solutions to identify suicidal individuals and ensure they receive life-saving resources are urgently needed. This study suggests that identifying at-risk individuals in need of such resources through social media may increase the likelihood of reaching at-risk people in real-time.

Our best-performing model performed well across most metrics and used text from posts' title and body as well as metadata (e.g. post type). This model outperformed one that used only text from post's title and body and excluded metadata, suggesting that inclusion of more information beyond the post's text is key to improving model performance. Other recent work has highlighted the importance of metadata as well, finding that lower average of mean 'likes' predicted suicidal posts on Instagram (Lekkas et al., 2021), suggesting suicidal content receives less engagement on social media This is similar to the user engagement findings in our study, given we found that users who made STB posts tended to have lower reputation scores than non-STB users (i.e. a RallyPoint specific-metric of user engagement and influence on the site). Future studies could consider including additional information about users such as features that may be available via the social media platform (e.g. age, gender).

In evaluating our final model, we prioritized sensitivity over precision because our goal was to be overly inclusive and detect more STB posts at the cost of making more false positive predictions. Low precision is common when trying to predict low base rate behaviors like STB. Although some have criticized the value of suicide risk models with low precision (Belsher et al., 2019), others have argued that even with low precision, suicide prediction tools can still have clinical value depending on how they will be used practically (Kessler, 2019; Kessler, Bossarte, Luedtke, Zaslavsky, and Zubizarreta, 2020). In the current project, the potential benefits of our risk algorithm outweigh low-precision issues because this model will be used to identify users for a series of low-cost, low-burden interventions delivered on the RallyPoint site. For example, we are using this model to identify users for a Barrier Reduction Intervention, which was adapted from other work by our group (Jaroszewski, Morris, & Nock, 2019). This is a psychoeducational intervention aimed at increasing users' likelihood of reaching out to mental health resources (e.g. Veteran's Crisis Line) if they need help (see Zuromski and Nock, 2023 for clinical trial registration).In this specific context, flagging a post as suicidal when the post is not suicidal does not pose a serious burden to the user; if the resource recommendation is not applicable, users can simply scroll past this suggestion. Missing a suicidal post is much more important and therefore high sensitivity was prioritized in our study. In contrast, if this model identified individuals for a more intensive intervention (e.g. psychiatric hospitalization), higher precision would be desired to ensure that the potential benefits of using the risk tool outweigh the costs, both financially and in terms of the intrusiveness of the intervention for an individual. As such, although we concluded in our study that the benefits of reducing false negatives outweighed the costs of more false positives, the risks of using a lower precision model to inform delivery of interventions should be weighed carefully on a case-by-case basis. This is especially important when the planned interventions are intensive (e.g. hospitalizations, welfare checks) and may pose risks such as invasion of privacy and risk of harm to the individual. In these cases, prioritizing higher precision over sensitivity may be desirable.

In addition to detecting posts with suicidal content, we were interested in examining these posts to better understand how Servicemembers and Veterans talk about suicide on social media. Little prior research has examined how military personnel talk about STB online, and this research has taken a retrospective approach, analyzing social media posts of Servicemembers who had died by suicide to identify any patterns in content leading up to the death (Bryan et al., 2018). We found key differences in content between STB and non-STB posts, which may help improve our ability to prospectively identify at-risk users. Specifically, STB posts were more likely to contain explicit suicidal language and were more likely to use first-person pronouns, which is a common language feature of individuals experiencing and expressing their own negative affect (Berry-Blunt, Holtzman, Donnellan, & Mehl, 2021). The higher usage of first-personal pronouns in suicidal posts is consistent with findings from other social media websites like Reddit (e.g. De Choudhury, Kiciman, Dredze, Coppersmith, & Kumar, 2016) and Twitter (O'Dea et al., 2015). In addition, STB posts were more likely to contain negations, which have been found to

https://doi.org/10.1017/S0033291724001557 Published online by Cambridge University Press

characterize posts in forums about suicide and depression on Reddit (Low et al., 2020), and the adverb 'just,' which communicates emphasis, immediacy, and simplicity (e.g. 'I just want to die') and has been found as a top word within suicidal text in a prior study (Franz, Nook, Mair, & Nock, 2020). We also compared metadata features between STB and non-STB posts, finding that the most common post type for both STB and non-STB posts were comments. This suggests that the type of social media interaction matters. For instance, only 4% of STB posts occurred in 'status updates,' or unique posts made by an individual user, whereas 91% of STB posts were comments made in response to other users. Thus, if we had only focused on users' own posts to develop our risk algorithm, we would be missing critical information on their suicide risk.

In addition, compared to the general RallyPoint user base, users in our sample were more likely to be older Veterans, for users making both STB and non-STB posts. Given we oversampled posts that were more likely to contain suicidal content, many of the non-STB posts contained mental health (but not suicide) related content. As such, the differences between our sample and the RallyPoint user base suggest that Veterans may be more interested or willing than Servicemembers to discuss sensitive topics such as mental health and suicide risk on social media. It may be that Veterans are more willing to self-disclose because they are not facing the same potential military repercussions as Servicemembers. STB disclosures while in service may impact Servicemembers' career and fitness to serve (e.g. access to weapons, security clearances). These types of impacts have been shown to reduce suicidal Servicemembers' willingness to seek treatment (Adler et al., 2020; VanSickle et al., 2016; Zuromski et al., 2019) and may also affect their willingness to share on social media. In terms of practical application, as described above, our model will be used to identify users for a light-touch psychoeducational intervention to encourage users to reach out to mental health resources. This intervention will contain relevant information for both Veterans and Servicemembers, so although posts used for model development were mostly from Veterans, we have sought to make our intervention broadly applicable and useful for all military personnel on the site.

Limitations

Our results should be interpreted in the context of several limitations. First, we observed a low incidence of STB posts on the RallyPoint site, which resulted in a class imbalance. This type of class imbalance is common in studies examining STB and reflects the low base rate of these behaviors, but nevertheless may create challenges for models. We addressed this using class weights in our final RoBERTa model. This approach was successful for our goals, although alternative balancing methods such as SMOTE or undersampling could be tried. In addition, because of the low incidence of STB posts, we were unable to conduct analyses on any demographic subgroups (e.g. posts only from Veterans or women). Second, although a strength of this study is that human labelers manually reviewed posts to code for the presence/absence of STB, a limitation is that the definition of STB was broad. That is, posts were coded as STB whether the user was describing current experience of STB or a past episode (e.g. suicide attempt in youth). Posts that contained descriptions of suicidal thoughts were coded the same as posts describing suicide plans or suicide attempt. For the purposes of identifying users who would benefit from immediate intervention on RallyPoint, refining the model to focus exclusively on descriptions of current STB and the presence of high suicidal intent (e.g. mentioning suicide plan, description of imminent suicide attempt) may be helpful. However, given the low incidence of STB posts, this granular coding was not possible in the current study. Third, prospective evaluation of the model is needed to determine how well the model identifies STB posts in real-time when deployed on the RallyPoint site. Fourth, because online language and the user bases of social media sites evolve over time, the model we developed here will likely need to be regularly updated and re-validated to continue to be useful and relevant for RallyPoint users. Lastly, although the RallyPoint user base is comprised of millions of U.S. military personnel that is demographically similar to the general U.S. military, these users may not be representative of Servicemembers and Veterans in general, which should be taken into consideration when generalizing findings to these populations.

Conclusions

These findings add to a growing literature suggesting the utility of harnessing social media to identify individuals who may be at risk for suicide. Further, we have filled a gap in this literature by focusing on military personnel, who are a group at particularly elevated risk for suicide. Given suicidal Servicemembers and Veterans may be hesitant to seek out treatment when they are struggling, our findings demonstrate the promise of using machine learning methods to identify potentially at-risk individuals outside of traditional health care settings. Methodologically, we show combining posts' text and metadata was needed to achieve high performance. Development of such risk models has important clinical applications in that they can be used to identify individuals who may benefit from receiving interventions and resources on social media. To that end, we currently are testing several real-time interventions that are delivered after a RallyPoint user is flagged by our risk model.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/S0033291724001557.

Code and data availability statement. All code is available at https://github.com/danielmlow/rallypoint_suicide_detection. Data cannot be shared given the sensitive nature of the content.

Acknowledgements. The authors wish to thank Pete Franz, Adam Jaroszewski, Erik Nook, LaTashia Raymond, Lilly Scherban, Chelsey Wilks, and Ellen Wittler who assisted with codebook development and data labeling.

Funding statement. This work was supported by an U.S. Air Force Research Laboratory AFWERX Small Business Innovation Research (SBIR) Grant (PIs: Dave Gowel & Matthew Nock; Grant number FA8649-20-9-9128). Dr Zuromski was supported by funding from the National Institute of Mental Health (K23MH120439; PI). Daniel M. Low was supported by a National Institute on Deafness and Other Communication Disorders training grant (5T32DC000038-28).

Competing interests. David Gowel, Carlos Madden, and Jonathan Epstein receive compensation in salary, consulting fees, and/or stock options from RallyPoint Networks, Inc. Dr Nock receives publication royalties from Macmillan, Pearson, and UpToDate. He has been a paid consultant in the past three years for Apple, Microsoft, and COMPASS Pathways, and for legal cases regarding a death by suicide. He has stock options in Cerebral Inc. He is an unpaid scientific advisor for Empatica, Koko, and TalkLife.

References

Adler, A., Jager-Hyman, S., Brown, G. K., Singh, T., Chaudhury, S., Ghahramanlou-Holloway, M., & Stanley, B. (2020). A qualitative investigation

of barriers to seeking treatment for suicidal thoughts and behaviors among army soldiers with a deployment history. *Archives of Suicide Research*, *24*, 251–268. https://doi.org/10.1080/13811118.2019.1624666

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2623–2631. https://doi.org/10.1145/3292500.3330701

Belsher, B. E., Smolenski, D. J., Pruitt, L. D., Bush, N. E., Beech, E. H., Workman, D. E., & …Skopp, N. A. (2019). Prediction models for suicide attempts and deaths: A systematic review and simulation. *JAMA Psychiatry*, *76*, 642–651. https://doi.org/10.1001/jamapsychiatry.2019.0174

Berry-Blunt, A. K., Holtzman, N. S., Donnellan, M. B., & Mehl, M. R. (2021). The story of "I" tracking: Psychological implications of self-referential language use. *Social and Personality Psychology Compass*, *15*, e12647. https://doi.org/10.1111/spc3.12647

Bryan, C. J., Butner, J. E., Sinclair, S., Bryan, A. B. O., Hesse, C. M., & Rose, A. E. (2018). Predictors of emerging suicide death among military personnel on social media networks. *Suicide and Life-Threatening Behavior*, *48*, 413–430. https://doi.org/10.1111/sltb.12370

Castillo-Sánchez, G., Marques, G., Dorronzoro, E., Rivera-Romero, O., Franco-Martín, M., & De la Torre-Díez, I. (2020). Suicide risk assessment using machine learning and social networks: A scoping review. *Journal of Medical Systems*, *44*, 1–15. https://doi.org/10.1007/s10916-020-01669-5

Colpe, L. J., Naifeh, J. A., Aliaga, P. A., Sampson, N. A., Heeringa, S. G., Stein, M. B., … Kessler, R. C. (2015). Mental health treatment among soldiers with current mental disorders in the army study to assess risk and resilience in service members (Army STARRS). *Military Medicine*, *180*, 1041–1051. https://doi.org/10.7205/MILMED-D-14-00686

De Choudhury, M., & De, S. (2014). Mental health discourse on reddit: Self-disclosure, social support, and anonymity. *Eighth International AAAI Conference on Weblogs and Social Media*, *8*, 71–80. https://doi.org/10.1609/icwsm.v8i1.14526

De Choudhury, M., Kiciman, E., Dredze, M., Coppersmith, G., & Kumar, M. (2016). Discovering shifts to suicidal ideation from mental health content in social media. Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, 2098–2110. https://doi.org/10.1145/2858036.2858207

Franz, P. J., Nook, E. C., Mair, P., & Nock, M. K. (2020). Using topic modeling to detect and describe self-injurious and related content on a large-scale digital platform. *Suicide and Life-Threatening Behavior*, *50*, 5–18. https://doi.org/10.1111/sltb.12569

Gamer, M., Lemon, J., Fellows, I., & Singh, P. (2019). irr: Various coefficients of interrater reliability and agreement. Retrieved from https://CRAN.R-project.org/package=irr

Gu, K., & Budhkar, A. (2021). A package for learning on tabular and text data with transformers. Proceedings of the Third Workshop on Multimodal Artificial Intelligence, 69–73. https://doi.org/10.18653/v1/2021.maiworkshop-1.10

Hoge, C. W., Castro, C. A., Messer, S. C., McGurk, D., Cotting, D. I., & Koffman, R. L. (2004). Combat duty in Iraq and Afghanistan, mental health problems, and barriers to care. *New England Journal of Medicine*, *351*, 13–22. https://doi.org/10.1016/s0084-3970(08)70136-1

Homan, S., Gabi, M., Klee, N., Bachmann, S., Moser, A.-M., Duri, M., … Kleim, B. (2022). Linguistic features of suicidal thoughts and behaviors: A systematic review. *Clinical Psychology Review*, *95*, 102161. https://doi.org/10.1016/j.cpr.2022.102161

Hughes, J. (2021). krippendorffsalpha: An R package for measuring agreement using Krippendorff's Alpha coefficient. *The R Journal*, *13*, 413–425. https://doi.org/10.32614/rj-2021-046

Jaroszewski, A. C., Morris, R. R., & Nock, M. K. (2019). Randomized controlled trial of an online machine learning-driven risk assessment and intervention platform for increasing the use of crisis services. *Journal of Consulting and Clinical Psychology*, *87*, 370–379. https://doi.org/10.1037/ccp0000389

Jones, N., Jaques, N., Pataranutaporn, P., Ghandeharioun, A., & Picard, R. (2019). Analysis of Online Suicide Risk with Document Embeddings and Latent Dirichlet Allocation. 2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW), 1–5. https://doi.org/10.1109/ACIIW.2019.8925077

Kang, H. K., Bullman, T. A., Smolenski, D. J., Skopp, N. A., Gahm, G. A., & Reger, M. A. (2015). Suicide risk among 1.3 million veterans who were on active duty during the Iraq and Afghanistan wars. *Annals of Epidemiology*, *25*, 96–100. https://doi.org/10.1016/j.annepidem.2014.11.020

Kessler, J. (2017). Scattertext: A browser-based tool for visualizing how corpora differ. Proceedings of ACL 2017, System Demonstrations, 85–90. https://doi.org/10.18653/v1/P17-4015

Kessler, R. C. (2019). Clinical epidemiological research on suicide-related behaviors—where we are and where we need to go. *JAMA Psychiatry*, *76*, 777–778. https://doi.org/10.1001/jamapsychiatry.2019.1238

Kessler, R. C., Bossarte, R. M., Luedtke, A., Zaslavsky, A. M., & Zubizarreta, J. R. (2020). Suicide prediction models: A critical review of recent research with recommendations for the way forward. *Molecular Psychiatry*, *25*, 168–179. https://doi.org/10.1038/s41380-019-0531-0

Lekkas, D., Klein, R. J., & Jacobson, N. C. (2021). Predicting acute suicidal ideation on Instagram using ensemble machine learning models. *Internet Interventions*, *25*, 100424. https://doi.org/10.1016/j.invent.2021.100424

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., & …Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. ArXiv. https://doi.org/10.48550/arXiv.1907.11692

Low, D., Rumker, L., Talkar, T., Torous, J., Cecchi, G., & Ghosh, S. S. (2020). Natural language processing reveals vulnerable mental health support groups and heightened health anxiety on Reddit during COVID-19: Observational study. *Journal of Medical Internet Research*, *22*, e22635. https://doi.org/10.2196/22635

Low, D., Zuromski, K., Kessler, D., Ghosh, S. S., Nock, M. K., & Dempsey, W. (2021). It's quality and quantity: The effect of the amount of comments on online suicidal posts. Proceedings of the First Workshop on Causal Inference and NLP, 95–103. https://doi.org/10.18653/v1/2021.cinlp-1.8

MacAvaney, S., Mittu, A., Coppersmith, G., Leintz, J., & Resnik, P. (2021). Community-level research on suicidality prediction in a secure environment: Overview of the CLPsych 2021 shared task. Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access, 70–80. https://doi.org/10.18653/v1/2021.clpsych-1.7

Monroe, B. L., Colaresi, M. P., & Quinn, K. M. (2008). Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, *16*, 372–403. https://doi.org/10.1093/pan/mpn018

Naslund, J. A., Bondre, A., Torous, J., & Aschbrenner, K. A. (2020). Social media and mental health: Benefits, risks, and opportunities for research and practice. *Journal of Technology in Behavioral Science*, *5*, 245–257. https://doi.org/10.1007/s41347-020-00134-x

O'Dea, B., Wan, S., Batterham, P. J., Calear, A. L., Paris, C., & Christensen, H. (2015). Detecting suicidality on Twitter. *Internet Interventions*, *2*, 183–188. https://doi.org/10.1016/j.invent.2015.03.005

Ophir, Y., Tikochinski, R., Asterhan, C. S. C., Sisso, I., & Reichart, R. (2020). Deep neural networks detect suicide risk from textual Facebook posts. *Scientific Reports*, *10*, 16685. https://doi.org/10.1038/s41598-020-73917-0

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, E. (2011). Scikit-learn: Machine learning in python. *The Journal of Machine Learning*, *12*, 2825–2830.

Ratner, A., Bach, S. H., Ehrenberg, H., Fries, J., Wu, S., & Ré, C. (2020). Snorkel: Rapid training data creation with weak supervision. *The VLDB Journal*, *29*, 709–730. https://doi.org/10.1007/s00778-019-00552-1

Ribeiro, J. D., Gutierrez, P. M., Joiner, T. E., Kessler, R. C., Petukhova, M. V., Sampson, N. A., & …Nock, M. K. (2017). Health care contact and suicide risk documentation prior to suicide death: Results from the Army Study to Assess Risk and Resilience in Servicemembers (Army STARRS). *Journal of Consulting and Clinical Psychology*, *85*, 403–408. https://doi.org/10.1037/ccp0000178

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, *10*, e0118432. https://doi.org/10.1371/journal.pone.0118432

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, *28*, 11–21. https://doi.org/10.1108/eb026526

https://doi.org/10.1017/S0033291724001557 Published online by Cambridge University Press

Suitt, T. H. (2021). *High suicide rates among United States service members and veterans of the post 9/11 wars*. Providence, RI: Watson Institute for International and Public Affairs, Brown University. Retrieved from https://watson.brown.edu/costsofwar/files/cow/imce/papers/2021/Suitt_Suicides_Costs%20of%20War_June%2021%202021.pdf.

Teo, A. R., Marsh, H. E., Liebow, S. B. L., Chen, J. I., Forsberg, C. W., Nicolaidis, C., & …Dobscha, S. K. (2018). Help-seeking on Facebook versus more traditional sources of help: Cross-sectional survey of military veterans. *Journal of Medical Internet Research*, *20*, e62. https://doi.org/10.2196/jmir.9007

VanSickle, M., Werbel, A., Perera, K., Pak, K., DeYoung, K., & Ghahramanlou-Holloway, M. (2016). Perceived barriers to seeking mental health care among United States Marine Corps noncommissioned officers serving as gatekeepers for suicide prevention. *Psychological Assessment*, *28*, 1020–1025. https://doi.org/10.1037/pas0000212

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., … Rush, A. (2020). Transformers: State-of-the-art natural language processing. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6

Zuromski, K. L., & Nock, M. K. (2023). Brief Interventions on Social Media to Reduce Suicide Risk (Intervention 3). ClinicalTrials.gov ID NCT06114849. Retrieved from https://clinicaltrials.gov/study/NCT06114849

Zuromski, K. L., Dempsey, C. L., Ng, T. H., Riggs-Donovan, C. A., Brent, D. A., Heeringa, S. G., & …Nock, M. K. (2019). Utilization of and barriers to treatment among suicide decedents: Results from the Army Study to Assess Risk and Resilience among Servicemembers (Army STARRS). *Journal of Consulting and Clinical Psychology*, *87*, 671–683. https://doi.org/10.1037/ccp0000400

https://doi.org/10.1017/S0033291724001557 Published online by Cambridge University Press