Original Paper

Analyzing Suicide Risk From Linguistic Features in Social Media: Evaluation Study

Cecilia Lao[1], BA, BACMP (Hons); Jo Lane[2], PhD; Hanna Suominen[1,3], MSc, MEdL, PhD, Docent (Adj/Professor)

[1]School of Computing, College of Engineering and Computer Science, The Australian National University, Canberra, ACT, Australia

[2]National Centre for Epidemiology and Population Health, College of Health and Medicine, The Australian National University, Canberra, ACT, Australia

[3]Department of Computing, Faculty of Technology, University of Turku, Turku, Finland

Corresponding Author:
Cecilia Lao, BA, BACMP (Hons)
School of Computing
College of Engineering and Computer Science
The Australian National University
145 Science Road
Canberra, ACT, 2600
Australia
Phone: 61 416236920
Email: cecilia.lao@anu.edu.au

Abstract

Background: Effective suicide risk assessments and interventions are vital for suicide prevention. Although assessing such risks is best done by health care professionals, people experiencing suicidal ideation may not seek help. Hence, machine learning (ML) and computational linguistics can provide analytical tools for understanding and analyzing risks. This, therefore, facilitates suicide intervention and prevention.

Objective: This study aims to explore, using statistical analyses and ML, whether computerized language analysis could be applied to assess and better understand a person's suicide risk on social media.

Methods: We used the University of Maryland Suicidality Dataset comprising text posts written by users (N=866) of mental health–related forums on Reddit. Each user was classified with a suicide risk rating (no, low, moderate, or severe) by either medical experts or crowdsourced annotators, denoting their estimated likelihood of dying by suicide. In language analysis, the Linguistic Inquiry and Word Count lexicon assessed sentiment, thinking styles, and part of speech, whereas readability was explored using the *TextStat* library. The Mann-Whitney *U* test identified differences between at-risk (low, moderate, and severe risk) and no-risk users. Meanwhile, the Kruskal-Wallis test and Spearman correlation coefficient were used for granular analysis between risk levels and to identify redundancy, respectively. In the ML experiments, gradient boost, random forest, and support vector machine models were trained using 10-fold cross validation. The area under the receiver operator curve and F_1-score were the primary measures. Finally, permutation importance uncovered the features that contributed the most to each model's decision-making.

Results: Statistically significant differences (*P*<.05) were identified between the at-risk (671/866, 77.5%) and no-risk groups (195/866, 22.5%). This was true for both the crowd- and expert-annotated samples. Overall, at-risk users had higher median values for most variables (*authenticity*, *first-person pronouns*, and *negation*), with a notable exception of *clout*, which indicated that at-risk users were less likely to engage in social posturing. A high positive correlation (ρ>0.84) was present between the part of speech variables, which implied redundancy and demonstrated the utility of aggregate features. All ML models performed similarly in their area under the curve (0.66-0.68); however, the random forest and gradient boost models were noticeably better in their F_1-score (0.65 and 0.62) than the support vector machine (0.52). The features that contributed the most to the ML models were *authenticity*, *clout*, and *negative emotions*.

Conclusions: In summary, our statistical analyses found linguistic features associated with suicide risk, such as social posturing (eg, *authenticity* and *clout*), *first-person singular pronouns*, and *negation*. This increased our understanding of the behavioral and thought patterns of social media users and provided insights into the mechanisms behind ML models. We also demonstrated the applicative potential of ML in assisting health care professionals to assess and manage individuals experiencing suicide risk.

Introduction

Background

Suicide is one of the leading causes of death worldwide [1] and is an international public health problem. The World Health Organization estimates that approximately 800,000 people die because of suicide every year, and global targets to reduce suicide mortality are unlikely to be met [2].

Effective suicide risk assessment screening methods are key to reducing this preventable cause of death [3,4]. Traditional approaches to suicide risk assessment include a comprehensive clinical evaluation and the use of self-reported measures, including the Columbia Suicide Severity Rating Scale, Patient Health Questionnaire, and other measures that screen for depression and psychological distress [1,5,6]. Although these approaches provide the best practice for suicide risk assessment, not all people experiencing thoughts of suicide or suicidal ideation disclose their risk or have access to health care professionals.

In addition, people experiencing suicide risk may not seek mental health support [7,8], and for those who do, the demand for clinicians often exceeds the supply, especially in remote areas where access to health care professionals is limited [9]. Therefore, an automated risk detection tool, or a deeper understanding of the linguistic features associated with suicide risk, could allow individuals to assess their own risk of suicide. This may prompt them to seek support and, in turn, increase suicide prevention.

Social Media and Suicide Risk Detection

Suicidal ideation has been widely documented on social media [10]. As these platforms provide individuals with an outlet to express their innermost thoughts [10], social media data offer new ways of understanding and assessing suicide risk. Hence, this creates novel possibilities for suicide assessment, intervention, and prevention [11].

Reddit, a web-based forum with >52 million daily users, offers particularly rich data. This is because of several reasons. First, it has a high character limit of 40,000 per post, which is a notable increase from other social media sites such as Twitter (280 characters), allowing users to write linguistically richer posts. Second, the website has the potential to be anonymous. Users can make *throwaway* accounts—temporary identities separate from their main accounts—to uninhibitedly discuss sensitive topics and emotions. This feature has been proven to promote open conversations and emotionally engaging feedback [12], thus making it ideal for suicide risk detection studies. Finally, Reddit's structure is advantageous. The website is made up of subforums (subreddits) that are topic specific. This allows researchers to preselect data from mental health–related subreddits, identifying users who potentially express suicide risk.

Machine Learning for Mental Health

In recent years, there has been increased interest in using machine learning (ML) to detect mental health conditions, including depression [13]. However, such studies often focus primarily on the performance of the classifier rather than on processes that underpin or explain its classification decisions [14].

This raises a key problem. ML models are often opaque, with *black box* models such as neural networks being largely uninterpretable [15]. This highlights a clear need for increased interpretability and understanding of the features themselves. Model-agnostic methods for understanding the feature importance include permutation importance [16] and Shapley Additive Explanations [17]. Such techniques are beneficial as they help us understand not only the outcomes but also the mechanisms behind the models themselves.

Research Objectives

This study aimed to examine the relationship between linguistic features and indicators of users' suicide risk on Reddit, thereby increasing interpretability. In addition to identifying statistically significant relationships, this study explored the contributions of the features to classifications by constructing ML models and permutation importance analysis.

Our main contributions are as follows: (1) we conducted nonparametric statistical analysis to identify linguistic features significantly associated with suicide risk; (2) we performed correlation analysis to identify relationships between significant features, thus identifying redundancies; (3) we built several ML models using linguistic features, highlighting the potential for future application; and (4) we measured the features that contributed the most to each model's decision-making through permutation importance analysis.

Methods

Data Selection and Access

In this work, we used the existing University of Maryland Suicidality Dataset [9,18]. This comprised social media posts annotated by mental health experts and crowdsourced annotators with respect to the author's suicide risk.

We chose this source for the following 3 main reasons.

First, it was extracted from the web-based Reddit forum. As stated earlier, Reddit has a generous character limit that allows greater linguistic complexity. Thus, it would be ideal to explore our first research question.

Second, another benefit of this data set was its high-quality annotations. A prevalent problem with social media data is the

reliability of ground truth labels; it is difficult to determine whether a web user is actually at risk in real life. Annotators are often inaccurate, even when label definitions are shown [19]. The Maryland data set alleviated this issue in several ways. To begin with, the researchers preselected at-risk (low-, moderate-, and severe-risk) users by identifying people who posted on mental health–related forums (eg, SuicideWatch). Furthermore, the annotation process was completed by mental health experts and crowdsourced annotators. Consensus mechanisms (eg, multiple annotators for each user) were also used.

Ethics Approval

The University of Maryland Suicidality Dataset [9,18] was approved for use by the Australian National University Human Research Ethics Committee (protocol number 2021/047). This was followed by obtaining proper permission to access and use it for the purposes of this study from the University of Maryland.

Data Overview

Reddit is a web-based forum designed to help people "detach from their real-world identities" [20]. The Maryland data set comprises text posts written by 934 unique users of this website—specifically, posts published on the SuicideWatch subforum from January 1, 2008, to August 31, 2015. It includes posts from both SuicideWatch and users' other non–mental health–related posts. In addition, users who did not post on any mental health–related forum [21] were included as a control group.

Although Reddit is intended to be anonymous, users may provide personal identifying information. Thus, this data set was further anonymized by replacing each username with a token, as well as by replacing all URLs [9,18].

Annotation Process

The Maryland researchers annotated the data set as follows. First, posts written by a given user were temporally organized and split into annotation units. These contained up to 5 posts each. Each unit was then annotated with a suicide risk rating by either medical experts or crowdsourced contributors. Experts were given *short* instructions asking them to follow their formal training in assessing patients at risk of suicide. Meanwhile, the crowdsourced annotators were given *long* instructions that asked

them to focus on risk factors such as thoughts (eg, suicide ideation and feeling like a burden), thought patterns (eg, sense of agitation), logistics (eg, talking about methods of attempting suicide), and context (eg, previous attempts and isolation from family and friends) [9,18].

Ratings were on a 4-point risk scale as follows [9]: (1) no risk ("I don't see evidence that this person is at risk for suicide"), (2) low risk ("There may be some factors here that could suggest risk, but I don't really think this person is at much risk of suicide"), (3) moderate risk ("I see indications that there could be a genuine risk of this person making a suicide attempt"), and (4) severe risk ("I believe this person is at high risk of attempting suicide in the near future").

Users with <10 posts and users whose posts had <3 annotators were eliminated from the data set by the Maryland researchers. This resulted in a final sample size of 866 unique users who posted on SuicideWatch, which is described by Reddit as a peer support forum for anyone struggling with suicidal thoughts. There was also an equal number of unannotated control users (n=866). Of the 866 annotated users, 245 (28.3%) were labeled by experts, whereas 621 (71.7%) were disjointly labeled by crowdsourced contributors.

The expert annotators included multiple mental health professionals. These included a cochair of the National Suicide Prevention Lifelines Standards, Training and Practices Sub-Committee, and a clinician in the Department of Emergency Psychiatry at Boston Children's Hospital [9,18]. To generate user-level annotations, maximum likelihood estimation was used [22,23]. Overall, the average Krippendorff interannotator agreeability α was .812.

In contrast, the crowdsourced task was completed on the web-based platform, CrowdFlower. The website's inbuilt consensus mechanism was used to resolve disagreements among crowdsourced annotations [18]. Each user was assigned a *trust score*, which indicated their reliability. Annotations were then weighted by this trust score and aggregated into a *confidence score* for each label. The label with the highest confidence score was chosen. This resulted in a Krippendorff α of .554.

An example of a typical post is presented in (Table 1). To preserve user privacy, the post body was an aggregate of several existing posts, and the subreddit was randomly chosen.

Table 1. Example of a typical Reddit post from the data set and the suicide rating.

Features	Value
Post ID	1a2b3c
User ID	45678
Time stamp (Unix epoch)	1.4E+09
Subreddit	r/self-harm
Post body	"I've been feeling depressed for a while. I don't know how to deal with it anymore..."
Label	Severe risk

Data Preprocessing and Linguistic Feature Engineering

In our research, we randomly split the data into an 80:20 training-test ratio following the Pareto principle [24]. This was
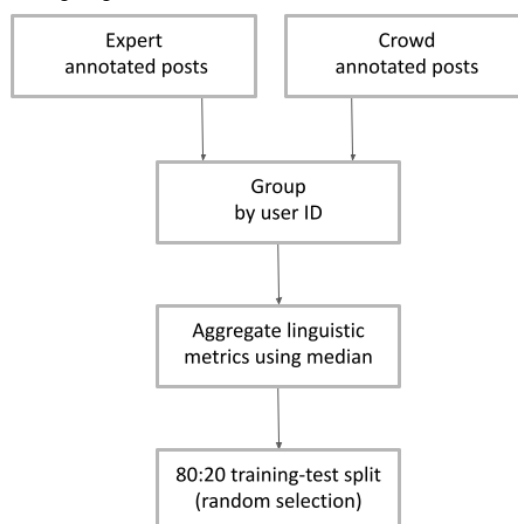
achieved by randomly selecting 80% of the user IDs from both the crowd- and expert-annotated data sets. All posts associated with the users were then retrieved. In addition, unannotated

control users and posts without any text were discarded. For the ML models, expert- and crowd-annotated users were combined into singular training and test sets to maximize the available data. Meanwhile, the statistical analysis was performed on the expert and crowd data sets separately to compare the distributions of the different groups.

Linguistic features for users were aggregated by taking the average of all the posts (Figure 1). The median rating was used instead of the mean rating to reduce the influence of outliers. Overall, we chose to group according to users to reflect the annotation process, as ratings were attached to a user rather than an individual post.

Linguistic Inquiry and Word Count (LIWC) 2015 and the *TextStat* Python library were used to extract linguistic features from posts. All the LIWC and *TextStat* features are listed in Multimedia Appendix 1 and Multimedia Appendix 2, respectively.

LIWC is a lexicon [25] that groups words into psychologically meaningful categories. Aside from aggregate features such as *authenticity*, the scores for most features were the percentage of total words in a text that belonged to a specific category. Prior studies have demonstrated the capacity of LIWC to detect emotionality [26,27], thinking styles [28], and individual differences [29,30]. Moreover, it has been used to detect self-reported symptoms of depression and other mental health conditions [31,32]. In this study, all the categories were used to ensure comprehensive coverage.

In juxtaposition, *TextStat* is a computerized analysis tool that measures linguistic complexity. This package was selected because it contains both simple features such as word count and widely used linguistic readability metrics such as the Gunning Fog Index, Simple Measure of Gobbledygook, and Flesch-Kincaid scores.

Figure 1. Flowchart detailing the data preprocessing stages.

Statistical and Correlation Analyses

Statistical and correlation analyses were designed and reported in consultation with the Australian National University Statistical Consultation Unit.

Statistical errors because of the assumption of normality are common in quantitative studies [33]. To mitigate this, we used the Shapiro-Wilk test [34] from *scipy.stats*. The analysis revealed that none of the features were normally distributed. Hence, nonparametric tests were used to compare the distributions of different risk groups.

First, a 2-sided Mann-Whitney *U* test was used to compare at-risk and no-risk users. This test was chosen because of its nonparametric nature and previous applications in medical studies [35]. To form this binary grouping, users who received either a severe-, moderate-, or low-risk rating were considered at risk. Meanwhile, users who received a no-risk rating formed their own group.

To supplement these results, we used the Kruskal-Wallis test. This compared the distribution of features within different risk levels. This analysis allowed us to determine whether the severe-, moderate-, and low-risk groups behaved differently.

For both tests, we used an α value of $P<.05$. To correct for multiple comparisons, we applied the Benjamini-Hochberg procedure [36] to the *P* values. We calculated 95% CIs to estimate the difference between medians using the Mann-Whitney *U* test. In addition, although post hoc methods (eg, Dunn test) for the Kruskal-Wallis test can be calculated to determine which specific medians are different, these were not computed in this work, as this was largely observed through the use of comparative box plots. Python (*scipy.stats* and *scipy.statsmodels*) and R (*wilcox* and *kruskal*) libraries were used for the implementation.

To reach a consensus between the expert- and crowd-annotated data sets, the features needed to have *P* values of <.05 and the same directionality to be labeled as *significant* in the Mann-Whitney *U* and Kruskal-Wallis tests. In addition, features with 95% CIs that included 0 were eliminated. This was because of 2 main reasons.

First, crowdsourced annotators were less reliable than experts. As they had less training, they had a lower macro F_1-score, with

a tendency to misclassify lower-risk users as having higher risk [9].

Second, the distribution of features could be different because of random variation. Although this does not necessarily mean that features that are only significant in one data set are not significant overall, it does suggest that the distributions are noticeably different. Therefore, it would be inappropriate to compare them.

A correlation analysis was also used to identify redundancies. This was because, in a practical context, having too many features limits interpretability and increases the computational complexity. For instance, if there are thousands of features, even if we know the weighted contribution of each feature, it is still extremely difficult to fully understand ML models and their classifications [37]. Hence, we identified relationships between significant features to determine potential proxies on both the expert- and crowd-annotated data sets, with P values corrected using the Benjamini-Hochberg procedure. We selected the Spearman correlation coefficient because of its nonparametric nature [38] and established use in medical research [39].

ML Models, Their Performance Evaluation, and Feature Importance Analysis

To determine whether the features would prove useful for risk assessment, we constructed several preliminary ML models that classified whether a user was at risk or had no risk. For this project, we used random forest (RF) [40], gradient boost (GB) [41], and support vector machines (SVMs) [42]. These techniques were selected because of their application in mental health research [43]. All LIWC and *TextStat* features were used to train the models.

An SVM is a supervised ML algorithm. It classifies data by representing each data point as a vector and fitting a hyperplane that separates the different classes [42]. In a 2D context, this is equivalent to fitting a dividing line through the data. Intuitively, an optimal hyperplane in such a fitting should be approximately at the center of the 2 classes. For SVMs, this is determined by calculating the distance between the hyperplane and the closest data points from each class. The hyperplane that maximizes this distance, or the maximum-margin hyperplane, is selected [42]. As not all data are linearly separable [44], SVMs use a kernel function in classification problems, a mathematical operation that performs the equivalent of mapping a lower-dimensional space to a higher dimension [45]. Ideally, this higher-dimensional projection should help make the data separable and, therefore, classifiable.

Decision trees are nonparametric and supervised learning methods. They work by splitting the root node, which represents the entire data set, into branch-like segments based on the values of their features. This continues until all the data are matched to a leaf node, which represents a class label [46]. Splitting is determined by the *purity* of the split, which is measured by metrics such as the information gain, gini index, and gain ratio [47].

Fundamentally, the algorithm tries to split the data so that the data points in each branch belong to the same class. However,

a problem with decision trees is that they are prone to overfitting [48]. Hence, a common way of addressing this problem is through ensemble methods that combine multiple smaller classifiers into a single classifier [49].

The RF method is a prime example of an ensemble method. It works by drawing k random subsamples of the data and fitting decision trees to each subsample. When presented with a new data point, each of the k decision trees casts a vote for the class label. The final label is determined by the results of the majority vote [40].

GB is another decision tree–based ensemble method. However, in contrast to RF, GB functions in an additive manner [50]. Fundamentally, this implies that each of the k decision trees is iteratively trained. The first decision tree is fitted to the training data, and the error is calculated. Following this, data points that were incorrectly classified will be given a higher weight, so that the following model can address the deficiencies of the previous model [41]. After all the weak learners have been trained, the final class label is determined by a weighted majority vote, with votes from more successful learners being more important.

As noted above, an 80:20 training-test split was used. The class distribution was as follows. In the training set, there were 63% (546/866) at-risk and 16.4% (142/866) no-risk users. A similar distribution was observed in the test set, with 14.4% (125/866) at-risk and 6.1% (53/866) no-risk users. To find the optimal hyperparameters and reduce overfitting [51], we used 10-fold cross validation [52] on the training set. The area under the receiving operator curve (AUC) was used as the primary scoring metric for validation because of increased discrimination and consistency [53,54]. To evaluate the performance on the test set, a more diverse range of metrics, including the AUC and accuracy, as well as the precision, recall, and F_1-score, were used to balance the trade-off between sensitivity and specificity [55]. Finally, confusion matrices [56] provided visualizations of true and false positives, as well as negatives on the test set.

Although univariate statistical tests can uncover relationships between linguistic variables and suicide risk, they might not indicate the importance of features in a given ML model. Hence, to better understand our models' decision-making process, we analyzed the permutation importance of each feature [16]. This examined the decrease in an existing model's score over a given number of iterations when the values of a single feature were randomly reordered. We implemented this using the Python *sklearn* library.

Features were considered important for a given model if a large decrease was observed and vice versa. For the purposes of this research, we calculated the permutation importance over 100 iterations on a holdout test set and used AUC, precision, and recall as the scoring mechanisms. Only variables with mean permutation importance values >1 SD away from 0 were considered significant.

XSL•FO

RenderX

Results

Mann-Whitney U Test

At-risk (low-, moderate-, and severe-risk) users had, on average, a greater use of *authenticity*, *first-person singular pronouns*, and *negation* (Multimedia Appendices 3 and 4). They also had lower *clout* (Tables 2 and 3). This suggests that, overall, they were more authentic in their expression and engaged less in social posturing (*authenticity* and *clout*). For brevity, all tables show only statistically significant values (*P*<.05) after applying the Benjamini-Hochberg correction, with the mean and CIs rounded to 4 significant figures.

This observation was also reflected visually. The box plots, which show that the overall distributions, in addition to the central measures such as the mean and median, were skewered further left for the at-risk users (Figures 2 and 3). Again, the inverse was observed for clout (Figures 4 and 5).

Table 2. Mann-Whitney U test results for expert-annotated users.

Feature	Examples	At-risk, mean (SD)	No-risk, mean (SD)	*P* value	95% CIs for differences between medians
Clout	N/A[a]	36.81 (16.62)	48.21 (11.48)	.005	−17.83 to −6.590
Authenticity	N/A	64.82 (21.31)	47.35 (20.86)	.005	10.04 to 27.09
First-person singular pronouns	I, my, and mine	7.105 (2.979)	5.419 (2.194)	.04	0.6900 to 2.840
Negation	Not, no, and never	1.391 (0.8524)	0.7924 (0.6987)	.01	0.2250 to 0.9650

[a]N/A: not applicable.

Table 3. Mann-Whitney U test results for crowd-annotated users.

Feature	Examples	At-risk, mean (SD)	No-risk, mean (SD)	*P* value	95% CIs for differences between medians
Clout	N/A[a]	32.00 (15.71)	40.48 (16.42)	<.001	−12.29 to −5.315
Authenticity	N/A	71.66 (19.57)	58.73 (20.18)	<.001	9.985 to 17.75
First-person singular pronouns	I, my, and mine	8.346 (2.902)	6.738 (2.579)	<.001	1.120 to 2.195
Negation	Not, no, and never	1.717 (1.072)	1.284 (1.031)	.001	0.1500 to 0.6000

[a]N/A: not applicable.

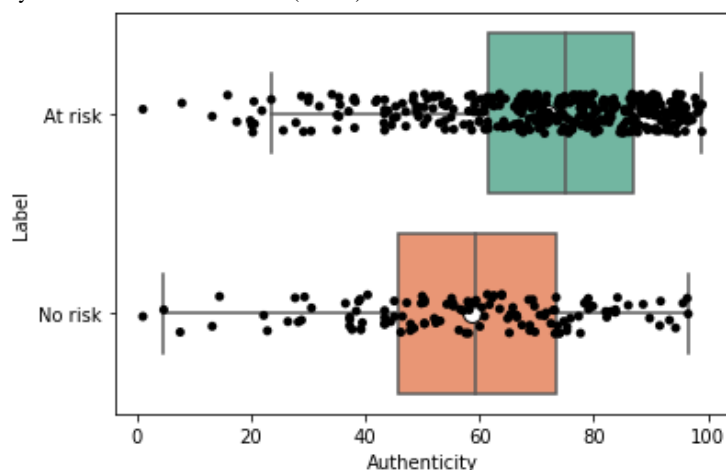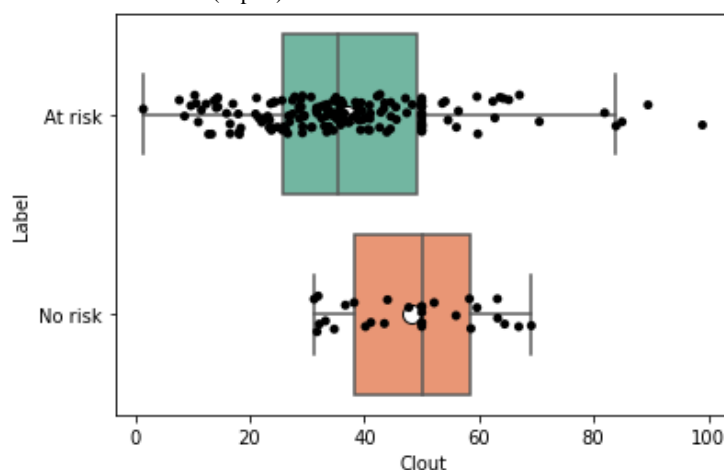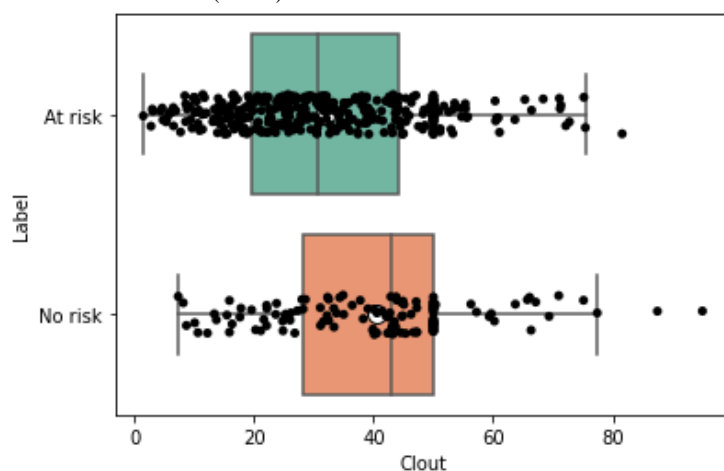**Figure 2.** Box plot for authenticity for at-risk and no-risk users (expert).
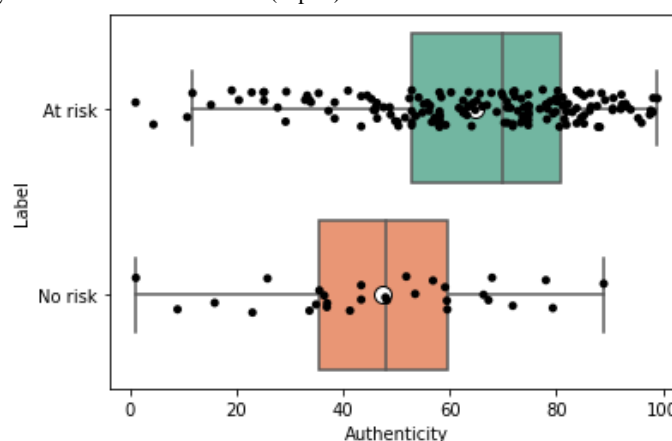
Figure 3. Box plot for authenticity for at-risk and no-risk users (crowd).

Figure 4. Box plot for clout for at-risk and no-risk users (expert).

Figure 5. Box plot for clout for at-risk and no-risk users (crowd).

Kruskal-Wallis Test

When comparing the severe-, moderate-, low-, and no-risk groups (Multimedia Appendices 5 and 6), no LIWC or *TextStat* features were significant (*P*<.05) for both the expert- and crowd-annotated data sets after correction. This indicates that although certain linguistic variables are associated with at-risk versus no-risk groups, there are no significant differences within the at-risk groups themselves.

Correlation Analysis

Using the Spearman correlation coefficient (Multimedia Appendices 7-10), we found the following.

For LIWC, various parts of speech (eg, *function words* and *pronouns*; ρ>0.74 and 0.84 [unless otherwise specified, the first ρ is the correlation coefficient on the crowd-annotated data set, whereas the second is for the expert-annotated data set; all

figures have been rounded to 2 significant figures]) were highly correlated with each other, indicating redundancy.

In addition, many parts of speech also had moderate correlations with variables measuring post length, such as syllable count (eg, *comparatives*; ρ>0.50 and 0.63) and word count (eg, *focus on the future*; ρ>0.59 and 0.67). This suggests that they could be proxies of length. Intuitively, this association was understandable, as the longer the post, the more parts of speech it will have.

Some features such as *clout* and *authenticity* appeared to be aggregate features, combining other variables, as shown through moderate positive correlations with other variables (eg, *authenticity* and *function words*; ρ>0.47 and 0.59). This was in line with the LIWC manual [57] and suggested that overall, aggregate features could be an efficient way of condensing information. Visual and statistical evidence provided additional support for this conclusion. When examining the box plots (Figures 2-5), we found that the aggregate measures were generally more discernible than individual categories. Furthermore, the *P* values tended to be smaller (*P*<.01), and the distance between medians also tended to be greater (Tables 2 and 3).

Model Performance

Overall, all models showed great promise in identifying suicide risk and achieved a similar performance (Table 3; Multimedia Appendices 11 and 12). Most errors lay in a tendency to overassign to the at-risk category. As type II errors are preferable to their type I counterparts in the medical domain, this demonstrated that LIWC and *TextStat* features are effective for building ML risk assessment models.

The performances of all models were largely comparable, with the AUCs for the GB, RF, and SVM models being 0.67, 0.66, and 0.68, respectively. Furthermore, when looking at performance evaluation outcomes (Table 4; Multimedia Appendices 11 and 12), all models were better at classifying at-risk users, with the RF model having the highest performance (F_1-score of 0.83) for this class. This was most likely because of the imbalanced nature of the data, with more users being at risk than not because of the selection process. It should be noted that the SVM, in particular, performed worse on the no-risk class, as indicated by its noticeably lower F_1-score (0.52). This implies that it is less useful in practice.

Table 4. Summary of classification results of various machine learning models[a].

Models	AUC[b]	Accuracy	Precision	Recall	F_1-score
Gradient boost	0.67	0.62	0.61	0.67	0.62
Random forest	0.66	0.75	0.65	0.66	0.65
Support vector machine	0.68	0.53	0.64	0.68	0.52

[a]The precision, recall, and F_1-scores are the macroaverage of the different classes.

[b]AUC: area under the receiving operator curve.

Permutation Importance

A noticeable overlap was present in the features that had higher permutation importance for the GB and RF models (Table 5), with *authenticity*, *negative emotion*, and *clout* contributing to higher precision and AUC (*authenticity* and *negative emotion* only) for both models. Meanwhile, the SVM yielded different results, having no common significant features with other models. However, as noted earlier, the SVM model had a notably lower performance (F_1-score) than the other 2 models. The permutation importance only measures the importance of a feature for a given model. Hence, if a model did not perform well, its permutation importance analysis results were not necessarily reliable. Thus, rather than showing that the aforementioned features were not important, this disparity could be an indicator of model quality.

Table 5. Permutation importance results for AUC[a], precision, and recall.

Features	Gradient boost, mean (SD)	Random forest, mean (SD)	Support vector machine, mean (SD)
AUC			
Authenticity	0.071 (0.041)	0.041 (0.027)	N/A[b]
Negative emotion	0.034 (0.024)	0.017 (0.01)	N/A
Clout	0.02 (0.016)	N/A	N/A
Whitespace	N/A	N/A	0.01 (0.005)
Precision			
Authenticity	0.057 (0.026)	0.030 (0.013)	N/A
Clout	0.018 (0.013)	0.035 (0.012)	N/A
Negative emotion	0.016 (0.014)	0.020 (0.011)	N/A
First-person singular pronouns	N/A	0.015 (0.008)	N/A
Quantitative processes	N/A	N/A	0.014 (0.010)
Informality	N/A	N/A	0.011 (0.008)
Recall			
Negative emotion	N/A	0.022 (0.015)	N/A
Positive emotion	N/A	0.021 (0.011)	N/A
Question mark	N/A	0.016 (0.007)	N/A
Affect	N/A	0.013 (0.008)	N/A
Function words	N/A	0.013 (0.008)	N/A
Colon	N/A	N/A	0.011 (0.004)
Ingest	N/A	N/A	0.01 (0.005)

[a]AUC: area under the receiving operator curve.

[b]N/A: not applicable.

Discussion

Principal Findings and Prior Work

A key finding was that linguistic features were significantly (*P*<.05, 95% CIs) associated with suicide risk on social media. This was achieved using nonparametric statistical analysis. Significant variables included social (*authenticity* and *clout*) and grammatical (*first-person singular pronouns* and *negation*) features (Tables 2 and 3). This confirmed prior studies linking suicide risk and depression to the increased use of first-person pronouns [31,58,59].

In addition to complementing prior work [31,58,59], our contribution provided novelty by examining the directionality and distribution of features at a finer granularity. We found that at-risk users tended to be more authentic and less concerned about social posturing (Tables 2 and 3). Overall, at-risk users had a larger median value than no-risk users for most features. However, there was no real difference between the distributions of the significant variables for low-, moderate-, and severe-risk users.

Another notable finding was the identification of redundant features such as various parts of speech. Although numerous studies have examined the relationship between linguistic features and adverse mental health [58,59], few statistically

examined the correlation between the significant features themselves. Moreover, although there are mathematical methods [17,60] for determining feature importance, these techniques are not widely used in the health sciences. Hence, established methods such as the Spearman correlation coefficient may be easier for clinicians to interpret.

Through correlation analysis, we found moderate positive relationships between readability metrics (eg, *Gunning Fog Index*; ρ>0.77 and 0.76), parts of speech (eg, *comparatives*; ρ>0.50 and 0.63), and post length (eg, *syllable*, *word*, and *sentence count*). This indicates that the underlying feature, length, could potentially be used in favor of its proxies. Moreover, using aggregate variables such as *clout* and *authenticity* may further increase computational efficiency. Not only do they combine more detailed categories, but they may also be better at discerning risk levels because of the increased differences between medians.

Another contribution was the demonstration that linguistic features alone could be used to create effective ML models (GB, RF, and SVMs). After hyperparameter tuning, the models achieved commendable AUCs ranging from 0.66 to 0.68 and F_1-scores ranging from 0.52 to 0.65. This received at par, if not better, performance than other lexical feature–based models whose AUC and F_1-scores ranged from 0.51 to 0.75 [61] and from 0.20 to 0.32 [62], respectively. In addition, all models had

a markedly better F_1-score for the at-risk group (Multimedia Appendices 11 and 12). As failing to identify a person with high suicide risk could lead to loss of life, a more conservative model is advantageous for suicide prevention.

Finally, we used permutation importance to identify the features that contributed the most to each model's decision-making. Through this analysis, we found that *authenticity* and *negative emotion* contributed to higher AUC and precision scores for both the GB and RF models, whereas *clout* contributed to a higher precision for the models. This indicated that such features could potentially be important indicators of suicide risk.

Reproducibility

As we are not the data set owners, we will not be able to provide it upon request. Thus, all applications for data access should be directed at the University of Maryland, following their formal request protocol.

By nature, ML for mental health is a sensitive research area. Hence, the source code for our experiments and the parameters of the classifiers will be made available upon reasonable request, with a justification for the intended use. All code distribution will be under the Massachusetts Institute of Technology license.

Limitations

Our study had 4 primary limitations. First, the observational nature of the study should be noted. Owing to privacy concerns, the University of Maryland Suicidality Dataset does not have ground truth labels, and we were unable to confirm whether users labeled as at risk were in fact experiencing suicidal ideation. In addition, it should be acknowledged that a person experiencing suicidal ideation may not be at risk of suicide, and people on the SuicideWatch forum may be affected by suicide through a family member or friend and not be experiencing suicide risk themselves. However, these confounds are likely to have been mitigated by expert annotation and consensus mechanisms.

Second, another limitation was the granularity of the annotations. Annotations were attached to each user and not to each post. Hence, we did not know which posts were more important and used aggregated features to train the models. Therefore, the performance could have potentially been further improved with finer-grained annotations.

Our third main limitation was the use of only linguistic features to train the models. As demonstrated by prior work, behavioral and relational analyses may further improve automated screening for suicide risk [35,63]. However, having a production-ready model was not the aim of this study. Instead, we aimed to determine whether simpler interpretable models could be used to screen for suicide risk. This was done to ensure that the models were accessible to health care professionals. Hence, *black box* ML methods such as deep learning and nonlexical features were not considered.

Our final limitation was the use of permutation importance to indicate the feature importance. As previously stated, permutation importance indicates only the importance of a feature for a particular model. Hence, it is arguably limited by the effectiveness of the models.

Future Work

This study focused on highlighting the usefulness of linguistic features in constructing ML models. Hence, only lexical features were used. However, prior studies [35,63] indicate that features based on behavioral data and metadata can be used to enhance performance. Therefore, before deploying our model for production, the inclusion of a more varied range of features could be investigated. It would also be interesting to explore deep learning as this would help us evaluate whether latent variables could further increase performance.

The usefulness of our findings in practice and how they relate to suicide assessment, intervention, and prevention could also be examined. This can be done in two ways: (1) exploring the use of ML-based models to support risk assessment on social media sites themselves and (2) investigating the integration of our work into clinical practice.

With regard to existing interventions on social media, in March 2020, Reddit developed *Reddit Care Resources*—an initiative aimed at providing mental health resources to users at risk of suicide or self-harm [64-66]. This method operates in 2 ways. First, if a user searches for certain keywords (eg, "suicide" and "kill myself"), the first result displayed is a post indicating where to find mental health support (Figure 6). Second, users can confidentially report other users who they believe are at risk of suicide or self-harm, which then connects them to trained crisis counselors [64,65].

Although these changes mark an increasing awareness of mental health and suicide risk, these measures could still be improved. For instance, the list of keywords that triggers Reddit Care Resources is limited, with searches for "depression," "self-harm," or "anorexia" not prompting this intervention (Figure 7).

ML models, such as those used in this study, could help alleviate this problem. For example, Reddit could run such models on searches and posts, prompting Reddit Care Resources to pop up if a certain risk threshold is met. This would eliminate the need to constantly expand the mental health–related keyword list, as internet slang and neologisms (eg, "proana" for "pro-anorexia") can make it difficult to record every word related to mental health.

Examining how our work can be integrated into clinical practice would also be meaningful. Social media can offer an outlet for people to express opinions and thoughts that they may find difficult to express face to face [12]. Hence, analysis of such posts by a health care professional may allow for a deeper understanding of their clients if informed consent is granted. However, a problem is that directly reading such posts may result in an unintentional breach of confidentiality [67]. For instance, if a client shares web-based posts with a health care professional that includes self-harm or abuse, they may be required to report this as part of their duty of care and mandatory reporting obligations [67].

Using a combination of ML and linguistic features (eg, LIWC and *TextStat*), as demonstrated in this work, could help address this problem. Being very time poor, health care professionals do not have time to read through social media posts. Instead,

with consent, automated methods could provide a report that summarizes suicide risk and other clinically relevant information, including the affective (eg, *emotional tone*), cognitive (eg, *discrepancy* and *certainty*), and social aspects (eg, *clout* and *authenticity*) of posts. This preserves client privacy while using ML to extract important clinical information that can potentially enhance client engagement and care. Furthermore, coproduction approaches with mental health experts and people with lived experience of suicide risk would help identify user and system requirements. This, in turn, would facilitate the development of future software apps (eg, desktop and mobile).

A final future development would be to diversify the annotated data sets. The University of Maryland Suicidality Dataset was unique because of its expert annotation and heightened levels of reliability; however, it has some limitations. For example, the demographics of Reddit tend to skew toward young and male [68-70], which is not representative of the world's population. Hence, gathering a wider and more varied data set would increase the generalizability of our work. Moreover, it may be helpful to further increase the granularity of annotations. There are 2 main reasons for this. First, it would help us understand which text posts contributed the most to an annotator's decision. Second, it would allow us to examine the fluctuation of risk within an individual, as a person identified as at risk may no longer be at risk at another point in time. These additions would likely allow us to achieve more informed results.

Figure 6. Screenshot of Reddit Care Resources.
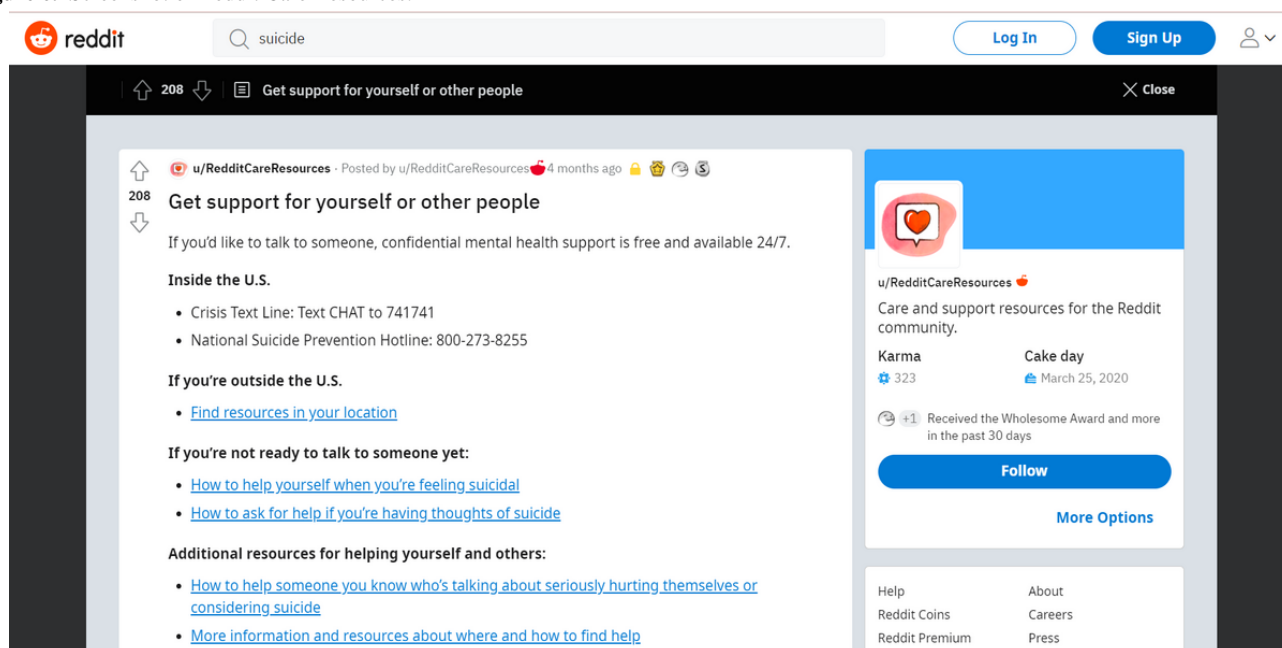
Figure 7. Screenshot of search results for "self-harm".
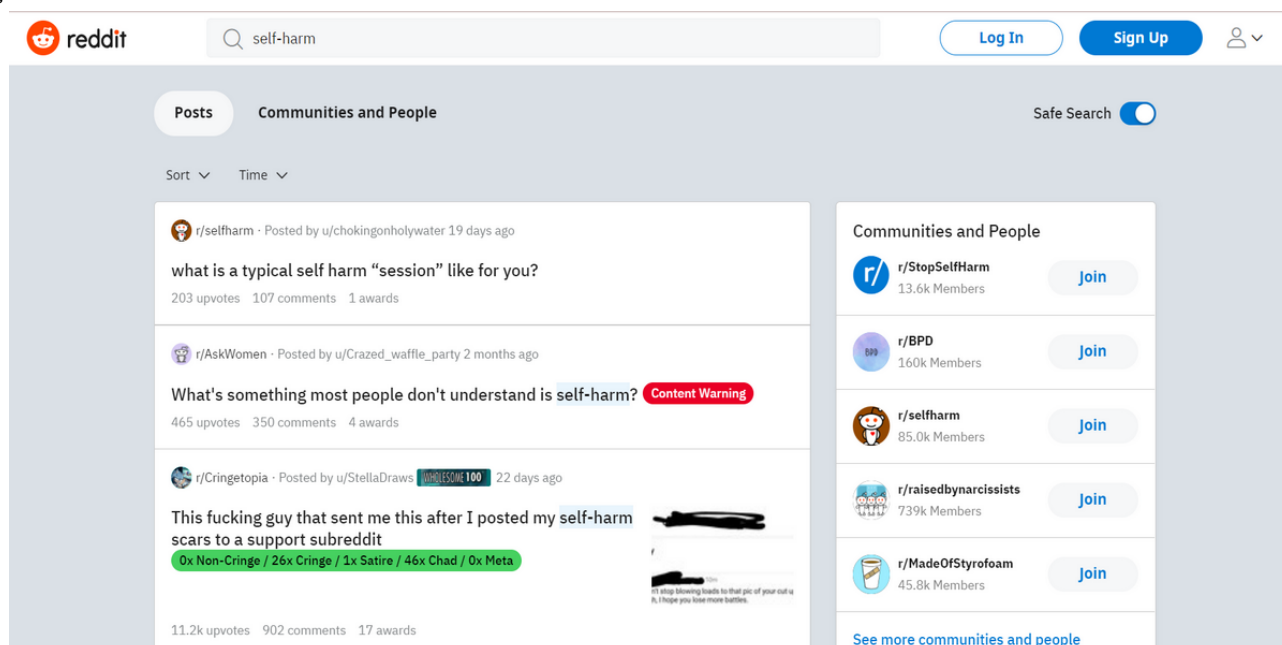
Conclusions

In this study, we demonstrated the potency of linguistic features in supporting suicide risk assessment through social media posts. Through statistical and permutation analyses, we were able to determine features significantly related to suicide risk, features that contributed the most to risk classifications, and redundancy through feature relationships. Finally, the commendable performances of the SVM, GB, and RF models highlight the utility of lexical features alone. This suggests that with future development, these models could be highly useful tools to help enhance clinical care.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Full glossary of Linguistic Inquiry and Word Count features.
[DOCX File , 28 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Full glossary of TextStat features.
[DOCX File , 20 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Full list of corrected *P* values for the Mann-Whitney U test (expert).
[XLSX File (Microsoft Excel File), 12 KB-Multimedia Appendix 3]

Multimedia Appendix 4

Full list of corrected *P* values for Mann-Whitney U test (crowd).
[XLSX File (Microsoft Excel File), 13 KB-Multimedia Appendix 4]

Multimedia Appendix 5

Full list of corrected *P* values for the Kruskal-Wallis test (expert).
[XLSX File (Microsoft Excel File), 12 KB-Multimedia Appendix 5]

Multimedia Appendix 6

Full list of corrected *P* values for the Kruskal-Wallis test (crowd).
[XLSX File (Microsoft Excel File), 13 KB-Multimedia Appendix 6]

Multimedia Appendix 7

Spearman correlation coefficients for Linguistic Inquiry and Word Count and TextStat features (expert).
[XLSX File (Microsoft Excel File), 159 KB-Multimedia Appendix 7]

Multimedia Appendix 8

Spearman correlation coefficient–corrected *P* values for Linguistic Inquiry and Word Count and TextStat features (expert).
[XLSX File (Microsoft Excel File), 162 KB-Multimedia Appendix 8]

Multimedia Appendix 9

Spearman correlation coefficients for Linguistic Inquiry and Word Count and TextStat features (crowd).
[XLSX File (Microsoft Excel File), 159 KB-Multimedia Appendix 9]

Multimedia Appendix 10

Spearman correlation coefficients–corrected *P* values for Linguistic Inquiry and Word Count and TextStat features (crowd).
[XLSX File (Microsoft Excel File), 165 KB-Multimedia Appendix 10]

Multimedia Appendix 11

Classification matrix for the at-risk class for machine learning models.
[DOCX File , 16 KB-Multimedia Appendix 11]

Multimedia Appendix 12

Classification matrix for the no-risk class for machine learning models.
[DOCX File , 16 KB-Multimedia Appendix 12]

References

1. Bolton JM, Gunnell D, Turecki G. Suicide risk assessment and intervention in people with mental illness. BMJ 2015 Nov 09;351:h4978. [doi: 10.1136/bmj.h4978] [Medline: 26552947]
2. Suicide in the world: global health estimates. World Health Organization. 2019. URL: https://apps.who.int/iris/handle/10665/326948 [accessed 2021-10-05]
3. Mann JJ, Apter A, Bertolote J, Beautrais A, Currier D, Haas A, et al. Suicide prevention strategies: a systematic review. JAMA 2005 Oct 26;294(16):2064-2074. [doi: 10.1001/jama.294.16.2064] [Medline: 16249421]
4. Zalsman G, Hawton K, Wasserman D, van Heeringen K, Arensman E, Sarchiapone M, et al. Suicide prevention strategies revisited: 10-year systematic review. Lancet Psychiatry 2016 Jul;3(7):646-659. [doi: 10.1016/S2215-0366(16)30030-X] [Medline: 27289303]
5. Brodsky BS, Spruch-Feiner A, Stanley B. The zero suicide model: applying evidence-based suicide prevention practices to clinical care. Front Psychiatry 2018 Feb 23;9:33 [FREE Full text] [doi: 10.3389/fpsyt.2018.00033] [Medline: 29527178]
6. Ryan EP, Oquendo MA. Suicide risk assessment and prevention: challenges and opportunities. Focus (Am Psychiatr Publ) 2020 Apr;18(2):88-99 [FREE Full text] [doi: 10.1176/appi.focus.20200011] [Medline: 33162846]
7. Essau CA. Frequency and patterns of mental health services utilization among adolescents with anxiety and depressive disorders. Depress Anxiety 2005;22(3):130-137. [doi: 10.1002/da.20115] [Medline: 16175563]
8. Rickwood DJ, Deane FP, Wilson CJ. When and how do young people seek professional help for mental health problems? Med J Aust 2007 Oct 01;187(S7):S35-S39. [doi: 10.5694/j.1326-5377.2007.tb01334.x] [Medline: 17908023]
9. Shing HC, Nair S, Zirikly A, Friedenberg M, Daumé III H, Resnik P. Expert, crowdsourced, and machine assessment of suicide risk via online postings. In: Proceedings of the 5th Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic. 2018 Presented at: CLPsych '18; June 5, 2018; New Orleans, LA, USA p. 25-36. [doi: 10.18653/v1/w18-0603]
10. Robinson J, Cox G, Bailey E, Hetrick S, Rodrigues M, Fisher S, et al. Social media and suicide prevention: a systematic review. Early Interv Psychiatry 2016 Apr;10(2):103-121. [doi: 10.1111/eip.12229] [Medline: 25702826]
11. Coppersmith G, Leary R, Crutchley P, Fine A. Natural language processing of social media as screening for suicide risk. Biomed Inform Insights 2018;10:1178222618792860 [FREE Full text] [doi: 10.1177/1178222618792860] [Medline: 30158822]
12. De Choudhury M, De S. Mental health discourse on reddit: self-disclosure, social support, and anonymity. Proc Int AAI Conf Web Soc Media 2014 May 16;8(1):71-80 [FREE Full text]
13. Linthicum KP, Schafer KM, Ribeiro JD. Machine learning in suicide science: applications and ethics. Behav Sci Law 2019 May;37(3):214-222. [doi: 10.1002/bsl.2392] [Medline: 30609102]
14. Velupillai S, Hadlaczky G, Baca-Garcia E, Gorrell G, Werbeloff N, Nguyen D, et al. Risk assessment tools and data-driven approaches for predicting and preventing suicidal behavior. Front Psychiatry 2019 Feb 13;10:36 [FREE Full text] [doi: 10.3389/fpsyt.2019.00036] [Medline: 30814958]
15. Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable AI: a review of machine learning interpretability methods. Entropy (Basel) 2020 Dec 25;23(1):18 [FREE Full text] [doi: 10.3390/e23010018] [Medline: 33375658]
16. Altmann A, Toloşi L, Sander O, Lengauer T. Permutation importance: a corrected feature importance measure. Bioinformatics 2010 May 15;26(10):1340-1347. [doi: 10.1093/bioinformatics/btq134] [Medline: 20385727]

XSL•FO
RenderX

17. Cohen SB, Dror G, Ruppin E. Feature selection based on the shapley value. In: Proceedings of the 19th International Joint Conference on Artificial Intelligence. 2005 Presented at: ICJAI '05; July 30-August 5, 2005; Edinburgh, UK p. 665 URL: https://www.ijcai.org/Proceedings/05/Papers/0763.pdf

18. Zirikly A, Resnik P, Uzuner Ö, Hollingshead K. CLPsych 2019 shared task: predicting the degree of suicide risk in Reddit posts. In: Proceedings of the 6th Workshop on Computational Linguistics and Clinical Psychology. 2019 Presented at: CLPsych '19; June 6, 2019; Minneapolis, MN, USA p. 24-33. [doi: 10.18653/v1/w19-3003]

19. Roß B, Rist M, Carbonell G, Cabrera B, Kurowsky N, Wojatzki M. Measuring the reliability of hate speech annotations: the case of the European refugee crisis. In: Proceedings of the 3rd Workshop on Natural Language Processing for Computer-Mediated Communication. 2016 Presented at: NLP4CMC III '16; September 22, 2016; Bochum, Germany p. 6-9 URL: https://www.linguistics.ruhr-uni-bochum.de/bla/ [doi: 10.17185/duepublico/42132]

20. Gutman-Wei R. Reddit's Case for Anonymity on the Internet. The Atlantic. 2018 Jun 29. URL: https://www.theatlantic.com/technology/archive/2018/06/reddit-anonymity-privacy-authenticity/564071/ [accessed 2021-10-05]

21. Pavalanathan U, De Choudhury M. Identity management and mental health discourse in social media. Proc Int World Wide Web Conf 2015 May;2015(Companion):315-321 [FREE Full text] [doi: 10.1145/2740908.2743049] [Medline: 27376158]

22. Dawid AP, Skene AM. Maximum likelihood estimation of observer error-rates using the EM algorithm. Appl Stat 1979;28(1):20-28 [FREE Full text] [doi: 10.2307/2346806]

23. Passonneau RJ, Carpenter B. The benefits of a model of annotation. Trans Assoc Comput Linguist 2014 Dec;2:311-326 [FREE Full text] [doi: 10.1162/tacl_a_00185]

24. Dunford R, Su Q, Tamang E. The Pareto principle. The Plymouth Student Scientist, University of Plymouth. 2014. URL: http://hdl.handle.net/10026.1/14054 [accessed 2021-10-05]

25. Tausczik YR, Pennebaker JW. The psychological meaning of words: LIWC and computerized text analysis methods. J Lang Soc Psychol 2009 Dec 08;29(1):24-54 [FREE Full text] [doi: 10.1177/0261927x09351676]

26. Blonder LX, Heilman KM, Ketterson T, Rosenbek J, Raymer A, Crosson B, et al. Affective facial and lexical expression in aprosodic versus aphasic stroke patients. J Int Neuropsychol Soc 2005 Oct;11(6):677-685. [doi: 10.1017/S1355617705050794] [Medline: 16248903]

27. Djikic M, Oatley K, Peterson JB. The bitter-sweet labor of emoting: the linguistic comparison of writers and physicists. Creat Res J 2006 Apr;18(2):191-197. [doi: 10.1207/s15326934crj1802_5]

28. Boals A, Klein K. Word use in emotional narratives about failed romantic relationships and subsequent mental health. J Lang Soc Psychol 2005;24(3):252-268 [FREE Full text] [doi: 10.1177/0261927x05278386]

29. Arguello J, Butler BS, Joyce E, Kraut R, Ling KS, Rosé C, et al. Talk to me: foundations for successful individual-group interactions in online communities. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 2006 Presented at: CHI '06; April 22-27, 2006; Montreal, Canada p. 959-968 URL: https://doi.org/10.1145/1124772.1124916 [doi: 10.1145/1124772.1124916]

30. Hancock JT, Landrigan C, Silver C. Expressing emotion in text-based communication. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 2007 Presented at: CHI '07; April 28-May 3, 2007; San Jose, CA, USA p. 929-932. [doi: 10.1145/1240624.1240764]

31. De Choudhury M, Gamon M, Counts S, Horvitz E. Predicting depression via social media. Proc Int AAI Conf Web Soc Media 2021 Aug 3;7(1):128-137 [FREE Full text]

32. Nguyen T, O'Dea B, Larsen M, Phung D, Venkatesh S, Christensen H. Using linguistic and topic analysis to classify sub-groups of online depression communities. Multimed Tools Appl 2017 Apr;76(8):10653-10676. [doi: 10.1007/s11042-015-3128-x]

33. Ghasemi A, Zahediasl S. Normality tests for statistical analysis: a guide for non-statisticians. Int J Endocrinol Metab 2012;10(2):486-489 [FREE Full text] [doi: 10.5812/ijem.3505] [Medline: 23843808]

34. Shapiro SS, Wilk MB. An analysis of variance test for normality (complete samples). Biometrika 1965 Dec 01;52(3/4):591-611. [doi: 10.2307/2333709]

35. Ramírez-Cifuentes D, Freire A, Baeza-Yates R, Puntí J, Medina-Bravo P, Velazquez D, et al. Detection of suicidal ideation on social media: multimodal, relational, and behavioral analysis. J Med Internet Res 2020 Jul 07;22(7):e17758 [FREE Full text] [doi: 10.2196/17758] [Medline: 32673256]

36. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B Stat Methodol 1995;57(1):289-300. [doi: 10.1111/j.2517-6161.1995.tb02031.x]

37. Ribeiro MT, Singh S, Guestrin C. Model-agnostic interpretability of machine learning. 2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016), New York, NY 2016.

38. Zar JH. Spearman rank correlation. In: Armitage P, Colton T, editors. Encyclopedia of Biostatistics. Hoboken, NJ, USA: John Wiley & Sons; 2005:7.

39. Mukaka MM. Statistics corner: a guide to appropriate use of correlation coefficient in medical research. Malawi Med J 2012 Sep;24(3):69-71 [FREE Full text] [Medline: 23638278]

40. Breiman L. Random forests. Mach Learn 2001;45:5-32 [FREE Full text] [doi: 10.1023/A:1010933404324]

41. Friedman JH. Stochastic gradient boosting. Comput Stat Data Analysis 2002 Feb;38(4):367-378 [FREE Full text] [doi: 10.1016/s0167-9473(01)00065-2]

42. Boser BE, Guyon IM, Vapnik VN. A training algorithm for optimal margin classifiers. In: Proceedings of the 5th annual workshop on Computational learning theory. 1992 Presented at: COLT '92; July 27-29, 1992; Pittsburgh, PA, USA p. 144-152 URL: https://doi.org/10.1145/130385.130401 [doi: 10.1145/130385.130401]

43. Cho G, Yim J, Choi Y, Ko J, Lee SH. Review of machine learning algorithms for diagnosing mental illness. Psychiatry Investigation 2019 Apr;16(4):262-269 [FREE Full text] [doi: 10.30773/pi.2018.12.21.2] [Medline: 30947496]

44. Elizondo D. The linear separability problem: some testing methods. IEEE Trans Neural Netw 2006 Mar;17(2):330-344. [doi: 10.1109/TNN.2005.860871] [Medline: 16566462]

45. Amari S, Wu S. Improving support vector machine classifiers by modifying kernel functions. Neural Netw 1999 Jul;12(6):783-789. [doi: 10.1016/s0893-6080(99)00032-5] [Medline: 12662656]

46. Song YY, Lu Y. Decision tree methods: applications for classification and prediction. Shanghai Arch Psychiatry 2015 Apr 25;27(2):130-135 [FREE Full text] [doi: 10.11919/j.issn.1002-0829.215044] [Medline: 26120265]

47. Patel N, Upadhyay S. Study of various decision tree pruning methods with their empirical comparison in WEKA. Int J Comput Appl 2012 Dec 18;60(12):20-25. [doi: 10.5120/9744-4304]

48. Li RH, Belford GG. Instability of decision tree classification algorithms. In: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. 2002 Presented at: KDD '02; July 23-26, 2002; Edmonton, Alberta, Canada p. 570-575. [doi: 10.1145/775047.775131]

49. Oshiro TM, Perez PS, Baranauskas JA. How many trees in a random forest? In: Proceedings of the 8th International Conference on Machine Learning and Data Mining in Pattern Recognition. 2012 Presented at: MLDM '12; July 13-20, 2012; Berlin, Germany p. 154-168. [doi: 10.1007/978-3-642-31537-4_13]

50. Chen T, Guestrin C. Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016 Presented at: KDD '16; August 13-17, 2016; San Francisco, CA, USA p. 785-794 URL: https://doi.org/10.1145/2939672.2939785 [doi: 10.1145/2939672.2939785]

51. Hastie T, Tibshirani R, Friedman J. Model assessment and selection. In: Hastie T, Friedman J, Tibshirani R, editors. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York, NY, USA: Springer; 2001:219-259.

52. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of the 14th international joint conference on Artificial intelligence. 1995 Presented at: IJCAI '95; August 20-25, 1995; Montreal, Canada p. 1137-1143. [doi: 10.5555/1643031.1643047]

53. Ling CX, Huang J, Zhang H. AUC: a statistically consistent and more discriminating measure than accuracy. In: Proceedings of the 18th international joint conference on Artificial intelligence. 2003 Presented at: IJCAI '03; August 9-15, 2003; Acapulco, Mexico p. 519-524. [doi: 10.5555/1630659.1630736]

54. Huang J, Ling CX. Using AUC and accuracy in evaluating learning algorithms. IEEE Trans Knowl Data Eng 2005 Mar;17(3):299-310 [FREE Full text] [doi: 10.1109/tkde.2005.50]

55. Buckland M, Gey F. The relationship between recall and precision. J Am Soc Inf Sci 1994 Jan;45(1):12-19 [FREE Full text] [doi: 10.1002/(sici)1097-4571(199401)45:1<12::aid-asi2>3.0.co;2-l]

56. Powers DM. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. J Mach Learn Technol 2020 Oct 11;2(1):37-63.

57. Pennebaker JW, Boyd RL, Jordan K, Blackburn K. The development and psychometric properties of LIWC2015. The University of Texas at Austin. 2015. URL: https://repositories.lib.utexas.edu/bitstream/handle/2152/31333/LIWC2015_LanguageManual.pdf?sequence=3&isAllowed=y [accessed 2021-10-05]

58. O'Dea B, Larsen ME, Batterham PJ, Calear AL, Christensen H. A linguistic analysis of suicide-related Twitter posts. Crisis 2017 Sep;38(5):319-329. [doi: 10.1027/0227-5910/a000443] [Medline: 28228065]

59. Rude S, Gortner EM, Pennebaker J. Language use of depressed and depression-vulnerable college students. Cogn Emot 2004 Dec;18(8):1121-1133. [doi: 10.1080/02699930441000030]

60. Sheikholeslami S. Ablation programming for machine learning. KTH Royal Institute of Technology. 2019 Aug 6. URL: https://www.diva-portal.org/smash/get/diva2:1349978/FULLTEXT01.pdf [accessed 2021-12-09]

61. Cheng Q, Li TM, Kwok CL, Zhu T, Yip PS. Assessing suicide risk and emotional distress in Chinese social media: a text mining and machine learning study. J Med Internet Res 2017 Jul 10;19(7):e243 [FREE Full text] [doi: 10.2196/jmir.7276] [Medline: 28694239]

62. Guan L, Hao B, Cheng Q, Yip P, Zhu T. Identifying Chinese microblog users with high suicide probability using Internet-based profile and linguistic features: classification model. JMIR Ment Health 2015 May 12;2(2):e17 [FREE Full text] [doi: 10.2196/mental.4227] [Medline: 26543921]

63. Calvo RA, Milne DN, Hussain MS, Christensen H. Natural language processing in mental health applications using non-clinical texts. Nat Lang Eng 2017 Jan 30;23(5):649-685 [FREE Full text] [doi: 10.1017/s1351324916000383]

64. Get support for yourself or other people. Reddit. 2021. URL: https://www.reddit.com/user/RedditCareResources/comments/own79d/get_support_for_yourself_or_other_people/ [accessed 2021-11-03]

65. What do I do if someone talks about seriously hurting themselves or is considering suicide? Reddit. 2020. URL: https://tinyurl.com/yc6ne78f [accessed 2021-11-03]

66.   What do I do if I'm thinking of seriously hurting myself or considering suicide? Reddit. 2020. URL: https://www.reddit
        help.com/hc/en-us/articles/360043075812-What-do-I-do-if-I-m-thinking-of-seriously-hurting-myself-or-consider ing-suicide-
        [accessed 2021-11-03]
67.   Jordan NA, Russell L, Afousi E, Chemel T, McVicker M, Robertson J, et al. The ethical use of social media in marriage
        and family therapy: recommendations and future directions. Fam J 2014;22(1):105-112 [FREE Full text] [doi:
        10.1177/1066480713505064]
68.   Distribution of Reddit Users worldwide as of July 2021, by gender. Statista Research Department. 2021. URL: https://www.
        statista.com/statistics/1255182/distribution-of-users-on-reddit-worldwide-gender/ [accessed 2021-10-05]
69.   Distribution of Reddit App users in the United States as of March 2021, by age group. Statista Research Department. 2021.
        URL: https://www.statista.com/statistics/1125159/reddit-us-app-users-age/ [accessed 2021-10-05]
70.   Percentage of internet users who use Reddit in the United Kingdom (UK) as of 3rd quarter 2020, by age group. Statista
        Research Department. 2020. URL: https://www.statista.com/statistics/1184024/reddit-user-demographics/ [accessed
        2021-10-05]

Abbreviations

AUC: area under the receiving operator curve
GB: gradient boost
LIWC: Linguistic Inquiry and Word Count
ML: machine learning
RF: random forest
SVM: support vector machine