Words of Suicide: Identifying Suicidal Risk in Written Communications

Amendra Shrestha, Nazar Akrami
Department of Psychology
Uppsala University
Uppsala, Sweden
firstname.lastname@psyk.uu.se

Lisa Kaati
Department of Information Technology
Uppsala University
Uppsala, Sweden
lisa.kaati@it.uu.se

Julia Kupper
Independent Researcher
Los Angeles, California, United States
info@juliakupper.com

Matthew R. Schumacher
Independent Researcher
Los Angeles, California, United States
mrschuma@lasd.org

2021 IEEE International Conference on Big Data (Big Data) | 978-1-6654-3902-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/BigData52589.2021.9671472

***Abstract*—Suicide is a global health problem with more than 700,000 individuals dying by self-destruction each year, yet it is classified as a low base rate behavior that is difficult to prognosticate. Aiming to advance suicide prediction and prevention, we examined the potential use of machine learning and text analyses models to predict suicide risk based on written communications. Specifically, we used a dataset consisting of more than 27,000 general writings unrelated to suicide, 193 genuine suicide notes from individuals who committed suicide, and an additional 89 suicide posts shared on sub-Reddits for an in-the-wild test to examine the prediction accuracy of two machine learning models (SVM & RoBERTa) and a linguistic marker model. Our tests showed that the machine learning models performed better than the linguistic marker model when examined on the test data. However, the linguistic marker model achieved higher results in the wild, correctly classifying 88% of written communications as a "high risk of suicide" versus 56% and 70% of the machine learning models. The best in-the-wild performing model was adopted in an online suicide risk assessment tool called Edwin to honor Edwin Shneidman for his numerous contributions to the field of suicidology. Finally, discrepancies between training and real-world data, vocabulary variation across domains, and the limited number of benchmarks constitute limitations that need to be addressed in future research.**

***Index Terms*—Suicide, machine learning, linguistic marker, RoBERTa, SVM**

I. Introduction

Self-destruction is a global health problem [1] with more than 700,000 people dying due to suicide each year, and many more suicide attempts not resulting in death (World Health Organization, 2021). However, lethal self-injury is considered a low base rate behavior that is difficult to predict, despite an improved understanding of recognizing factors that may facilitate in detecting high-risk individuals. There have been no significant advances in suicide prevention in the last forty years [2] and an effective algorithm that can anticipate self-inflicted deaths in a clinical setting is yet to be developed.

In combination with other corroborations, suicide letters are crucial evidence for investigating entities and can provide essential support in determining the manner, cause and/or circumstances of a death (e.g., [3] [4]). Thus, words can assist in classifying if a demise was self-inflicted and intentional (a "suicide") and also help identify potential motives for the self-induced annihilation because suicide notes are direct indicators of the author's decision and intent of self-destruction. As words precede actions, suicide letters are authored shortly prior to the act of lethal self-injury [5] and can be produced in different formats with varying styles: e.g., hand- and type-written documents or audio/video files. The deceased's reality is often expressed through final thoughts, concerns and feelings in an acknowledgement of "symbolic resolutions of tensions, problems, and failures that are embedded in the life dramas of suicidal individuals" [6], and can act as a means of communicating and connecting with those left behind.

The scientific study of self-induced annihilation, suicidology, began in 1957 when Shneidman and Farberow [7] compared a set of authentic suicide notes, written by individuals who had committed suicide, with letters authored by a control group of non-suicidal persons. This content study focused on the words and phrases used within these texts, and showed that the notes differed: genuine suicide letters contained more neutral statements than their fake counterparts - such as instructions for belongings - and more statements of deep discomfort - for instance hatred, demands, vengeance, self-blame and/or ascribing blame to survivors.

Discourse-based analyses have been an important component for risk assessments involving self-injured deaths. Nouns, verbs, adjectives and adverbs uttered by suicidal minds have been examined as code in attempting to encrypt the messages of lethal self-injury in clinical and research settings. As Shneidman [8] explains: "the proper language of suicidology is *lingua franca* - the ordinary everyday words that are found in the verbatim reports of beleaguered suicidal minds. It is the words that suicidal people say about their psychological pain and their frustrated psychological needs that make up the essential vocabulary of suicide". As individuals at risk of lethal self-injury often experience tunnel vision and constrictive thinking as a mental state [9], it inevitably effects their conceptuality and they produce a specific type of language

- that of an 'insider' [10]. In turn, suicidal tendencies are frequently expressed in common contextual themes and can range from hopelessness and loneliness, general negativity paired with unrealistic optimism, references to psychological and/or physical pain, perception of being a burden to bi-polarity (happiness and peacefulness versus grief, guilt or self-blame), method(s) used to attempt the suicide, and/or giving specific instructions for funerals, finances and/or final affairs.

With a shifting landscape to an ever-increasing online world, these written clues of emotions and feelings, but also intent of self-destruction, started progressing towards social media, online forums and personal websites in recent years; especially younger generations may be less likely to author a handwritten suicide letter or engage in face-to-face therapy sessions with a psychologist. Though traditional suicide notes were typically not distributed to a wider audience, due to the personal nature of the document, suicidal communications are more public than ever at present. Subsequently, the language use in online and offline communications can now be systematically analyzed and compared within a large dataset of suicide notes and control groups, which may assist in being able to predict and unravel some of the mysteries of self-inflicted deaths.

A. Outline and Aim

The aim of this work is to develop a tool that can assist in assessing the risk of suicide in written communications by combining forensic linguistic, psychological and computational techniques. Different approaches for risk assessment will be tested with the aim of developing a practicable suicide assessment tool that can identify if the author of a text might have a high risk of lethal self-injury by providing an acute snapshot of his/her language.

To detect suicide risk in written communications, a *linguistic indicator approach* and a *machine learning approach* with two different algorithms will be applied. These algorithms will be trained and tested on a dataset comprising texts from internet users that are (a) *non-suicidal*, (b) *users that are active on suicide discussion forums*, and from (c) *authentic suicide letters written by individuals who have committed suicide*.

The results will be implemented in a practicable tool: a digital assistant that can be used to determine linguistic risk indicators and assess if there is an imminent threat to the author's life based on his/her written communication. The tool could provide a flash warning of acute psychological distress or unendurable mental pain, which in turn could suggest that there is a suicide in progress. In honor of Edwin Shneidman and his numerous contributions to the field - and returning to his roots of assessing suicidal behavior based on language - the tool is called *Edwin*.

II. Background

While most suicide research is focused on offline settings where the individual is present and can answer questionnaires, more recent studies have addressed the problem of detecting suicide letters or detecting suicide ideation in written digital/online communications. One of the reasons for focusing on written communications is that our language - and the way we express ourselves - contains information that can be a valuable source for detecting and analyzing various aspects of self-destruction, such as suicidal ideation and other indicators that might affect the risk of suicide. Suicide risk can be defined as the likelihood that an individual will die by suicide, which in turn could be motivated by grievances, interpersonal or financial problems, and/or physical or psychological illnesses.

The American Association of Suicidology [11] published a list of several warning signs that can indicate an acute risk of suicide, which are considered the consensus. These include:

- Threatening or talking about wanting to hurt or kill oneself
- Identifying ways to kill oneself by seeking access to firearms, pills, or other means
- Talking or writing about death, dying or suicide, when these actions are out of the ordinary
- Increased substance use
- No reason for living or no sense of purpose in life
- Dramatic mood changes
- Being unable to sleep or sleeping all of the time

In addition, the U.S. National Institute of Mental Health [12] released a list of warning behaviors that may be signs of an individual thinking about suicide:

- Great guilt or shame, being a burden to others
- Feeling empty, hopeless, trapped, or having no reason to live, feeling extremely sad, anxious, agitated or full of (uncontrolled) rage
- Changing behavior and starting to make plans or researching ways to die
- Withdrawing from friends, family and society
- Saying good-bye, giving away important items or drafting a will

One approach to observe an increased risk of lethal self-injury is to detect suicidal ideation. This might - or might not - include a plan for committing suicide. Shaoxiong et al. [13] define suicidal ideation detection (SID) as determining whether a person has suicidal thoughts by analyzing tabular data of an individual or textual content written by that person. According to Lopez-Castroman et al. [14], online environments where people communicate and express their feelings, sufferings and thoughts, are a natural source for SID and mining social media might also be useful to improve suicide prevention. Ophir et al. [15] used Amazon's Mechanical Turk to obtain Facebook data from 1,002 subjects. The participants completed eight psycho-diagnostic measures and provided the researchers with twelve months of Facebook posts. Two different models were created: one that predicted suicide risk from those Facebook posts and a second one that combined the Facebook posts with the psycho-diagnostic measures to predict suicide risk. The model that operated on the combined data showed better

prediction accuracy compared to the model operating only on Facebook data; however, the researchers concluded that machine learning-based analyses of everyday social media activity can improve suicide risk predictions.

A different methodology to detect suicidal ideation using machine learning was presented by Sawhney et al. [16] in 2018: the researchers used data from Twitter that expressed suicidal ideation in words and phrases such as "suicide", "end my life", "wanna die" and "kill me now". The tweets were then marked as "suicidal intent present" or "suicidal intent absent" by three annotators; around 15% of the tweets were considered to contain suicidal intent. By using machine learning features from LIWC (Linguistic Inquiry and Word Count), as well as data dependent features, Pennebaker et al. [17] achieved an accuracy of 0.86 when classifying tweets containing suicidal intent in comparison to tweets without suicidal intent.

In contrast, Cheng, Chang & Yip [18] studied individuals that conversed about self-induced annihilation on the Chinese microblog Weibo. They used an online survey on Weibo users that compared differences in psychological and social demographic characteristics between those who engaged in suicidal communication and those who did not. The participants were assessed using six different measures: suicidal communication, suicidal ideation, negative affectivity, vulnerable personality and their preference for using social media and demographics. While the research did not consider any traits from the actual communications on Weibo, the results are still interesting: Cheng et al. concluded that greater suicide ideation, negative affectivity, neuroticism and lower agreeableness were found to be correlated with suicidal communications on Weibo.

When it comes to detecting the risk of self-destruction, researchers commonly use suicide notes to develop suitable methods. For instance, [19] used machine learning to distinguish between genuine and elicited suicide notes. In an experiment with letters from 33 suicide completers and 33 fabricated notes (adopted from Shneidman and Farberow, 1957), they found that the machine learning algorithm performed better that mental health professionals. Specifically, the algorithm correctly identified the suicide letter in 78 percent of the cases, whereas the mental health professionals only correctly identified suicide letters in 63 percent of the cases. In a different study of 286 suicide notes with additional findings from 33 real and 33 fabricated notes, Shapero [20] found that the language of genuine suicide notes included affections, future tense, references to family members, pronouns, names, negatives, intensifiers and maximum quantity terms.

III. Method

A. Two Approaches to Detect Suicide Communication

We used two approaches to detect communications with a high risk of lethal self-injury: a *linguistic indicator approach* and a *machine learning approach* with two different algorithms. These two methods differ since one of them is theory-driven and the other one is data-driven. A theory-driven approach takes its starting point in a theory; i.e., in this study the consensus indicators for suicide risk from the list of warning behaviors from the National Institute of Mental Health [12]. Our data-driven approach employed data to identify traits that separated suicidal risk from normal conversations. In addition, two different machine learning algorithms were utilized: a Support Vector Machine (SVM), which was used as benchmarking, and a neural network that was built on the Bidirectional Encoder Representations from Transformers (BERT) [21]. These two different approaches applied distinct features in the text to separate suicidal ideation and suicidal risk from normal conversations on the internet. To train and test the different approaches, data consisting of a set of texts from several different locations were selected. A high risk of suicide indicated that the individual likely authored a genuine suicide letter and that he/she had intentions to commit suicide.

B. Data

The data used to train and test our models consisted of a total of 27,329 texts, with the lengths of the different texts ranging from 50 to 20K characters. These lengths were varied to bring randomness and prevent biases based on the text sizes. As part of the preprocessing, all URLs and links were removed from the texts. To identify a high suicide risk, we analyzed a set of 193 genuine suicide notes compiled by Dr. John Olsson from the *Forensic Linguistics Intelligence*. Data from a wide range of social media platforms was used as a normal population, as well as data from a discussion forum called *The Suicide Project*, a support site and place where users can share their stories of suicide despair and hope with others. The reason for including data from *The Suicide Project* was to train our algorithms to differentiate between a suicide note and communications that may express suicidal despair or ideation. Table I lists sources of the data that was used in our experiments.

To test how well these approaches worked "in the wild", a dataset (89 posts) consisting of suicide posts from Reddit was utilized, collected by the Reddit user *u/IncelGraveyard*. The suicide posts are from incels that have been active on Reddit. Descriptive statistics about the authors of the posts were unavailable but it can be assumed that the majority of the authors are male and that they - to some extent - identify themselves as incels (for more information see [22]).

C. Linguistic Indicator Approach

Based on previous research on suicidal ideation (e.g., [3] [4] [12] [7]), we identified a set of indicators that can be used to detect suicidal ideation. To assess the indicators, information about them from a given text using a dictionary-based approach was extracted. Each dictionary contained a set of words that represented an indicator; we then counted the relative frequencies of the words in the text material. The prevalence of the dictionary words in the texts were thus standardized, divided by total word counts, and produced a score for each variable that represented its relative frequency of occurrences in the text. This gave an indication of the presence of each indicator in the target text. Some of the indicators had psychological characteristics and were latent

TABLE I
SOURCE OF THE DATA AND NUMBER OF TEXTS USED FOR TRAINING AND TESTING.

Source	Training (80%)	Test (20%)
Suicide notes (from individuals who committed suicide)	155	38
Suicide project (discussion forum)	1022	255
Boards (discussion forum)	3120	779
Daily Stormer (website)	1107	276
Gab (social media)	980	244
Gates of Vienna (blogs)	1062	265
Google blogs (blogs)	2710	677
Incels (discussion forum)	1210	302
Islamic awakening (discussion forum)	836	208
Lookism (discussion forum)	36	8
Looksmax (discussion forum)	789	197
Neogaf (discussion forum)	1752	437
Niggermania (discussion forum)	364	91
Reddit (discussion forum)	2954	738
Stormfront (discussion forum)	1765	441
Turn to islam (discussion forum)	1067	266
VNN forum (discussion forum)	943	235
Total	**21872**	**5457**

constructs with no absolute values; these types of values are only meaningful in relative terms. Subsequently, the scores for each indicator were compared to the scores of a normal population, which consisted of a set of texts from a wide range of forums and social media posts (see Table I).

When working with a dictionary-based analysis, it is important to consult experts with significant domain knowledge of the studied environment. Thus, we consulted a number of specialists to create a set of dictionaries for the indicators that were applied. The experts utilized the procedure described in [23]. A total of eleven dictionaries were created and used to extract eleven different scores for each text in our dataset - one score for each indicator (see Table II).

D. Machine Learning Approaches

We used two different machine learning models to detect suicide risk in written communications: SVM and RoBERTa (Robustly Optimized BERT Pretraining Approach). When training the SVM model, all texts were converted to lower case and hyper-parameter tuning was done utilizing grid search to estimate the optimal parameters of the classifier. To select features we used TF-IDF (Term Frequency–Inverse Document Frequency) numerical statistics that reflect how important a word is to a text in a collection of texts. English stop words - words that does not add much meaning to a sentence e.g., a, the, is, are - were removed from the text before applying TF-IDF.

While classical machine learning approaches, such as SVM, make use of a bag of words (BOW) to create numerical features, the sequential order of the words and their relations were not considered. RoBERTa is a transformer language model which captures the contextual relationship between words in text data [24]. We used RoBERTa, which is built on BERT with a modification of hyper-parameters and the removal of the next-sentence pre-training objective. RoBERTa has been trained with a large corpus, including news articles, to achieve a better performance in understanding natural language tasks.

Instead of training the model from scratch, we utilized a pre-trained RoBERTa model and fine-tuned it with our suicide letter data. The model consisted of a RoBERTa-base with a classifier layer on top. A RoBERTa tokenizer was applied, which has the rules to tokenize text, as well as the vocabulary and dictionary mapping tokens to numerical indices. The maximum sequence of token was fixed to 512 tokens, and the Adam optimizer was chosen. We are doing a fine-tuning of the weights and thus, a smaller learning rate of 5e-6 was set. Since we did a classification task to determine if a text contained suicidal risk or not, a spare categorical cross entropy was chosen as a loss function. The experiment was completed with 2 epochs and the batch size was kept at 8. During training process, we chose the best performing model measured by accuracy on the validation set. To train and test the two different machine learning models, the dataset described in Section III-B was used. 20% of texts were randomly selected as test data and not used in the training.

IV. RESULTS

A. Suicidal risk

As mentioned in the method section, we utilized 80% of the texts as training data and 20% as test data. While fine tuning RoBERTa model, 20% of training data is taken as validation data. We used all texts from the different sources in Table I as the normal (non-suicide) negative class and the suicide letters as the positive class. The trained SVM and RoBERTa models were used to classify the test data. The results showed that the SVM model correctly classified 89% of the suicide texts, whereas the corresponding scores for the RoBERTa model was 97% (see Table III).

For the *linguistic indicator approach*, using the training sample only, we started by creating a binary variable with the positive classes (suicide letters) as one group and the negative classes (normal texts) as the second group. Subsequently, a series of receiver operating characteristic (ROC) analyses were conducted to identify the discrimination threshold on each of

TABLE II
INDICATORS OF SUICIDE RISK.

Indicator	Description	Example words	AUC
Suicide	Expressions of ways to commit suicide	suicide, die, drown, drugs, alcohol	.67
Suicidal communication	Expressions related to reasons for suicide	depression, hell, alone, unbearable	.88
Existential communication	Expressions of existential anxiety	understand, fear, choice, guilt, god	.76
Social connections	Expressions of social connections/relations	friend, children, father, mother, family	.72
Personal pronoun	Expressing specific self-reference	he, she, I, we, they, us, themselves	.90
Grievance	Use of grievance terminology	disappointing, heartache, unfair, suffer	.33
Anxiety	Use of anxiety terminology	afraid, alarmed, doubt, danger, worry	.33
Negative emotions	Expressions of generally negative emotions	bash, careless, offend, whine, unsuccessful	.63
Positive emotions	Expressions of generally positive emotions	beauty, beloved, like, kiss, sweet	.71
Anger	Use of anger terminology	abuse, hate, asshole, kill, jerk	.29
Violence	Use of violence/power-related terminology	execution, terrorize, stab, smack, war	.30

the indicators for our binary variable (suicide letters vs. normal texts). The ROC scores varied between .29 and .90, and on average departed from chance (.50) by .23. Having identified the discrimination threshold for all 11 variables/dictionaries using the coordinates of the curve, we then generated 11 new variables where each text was assigned a zero (0) or one (1), depending on their scores on the original variable and relative to the threshold, being either above or below the threshold. Next, for each text we averaged the scores across the 11 new variables arriving at a new variable ranging between 0 and 1. To simplify matters, the average score was multiplied by 100, arriving at our final variable, which now ranged between 0 and 100. We denoted this variable as the Suicide Risk Score (SRS). Finally, we conducted a ROC analysis with our binary variable (suicide notes vs. normal texts) and the SRS. The results revealed that the area under the curve (AUC) was .99 (SE = .004, 95% CI [.980, .996], see Fig. 1).

More importantly, the discrimination threshold from the analyses above was utilized to classify the texts in the test data. The results of these analyses showed that the model correctly classified 90% of the suicide letters and 96% of the texts from the normal (non-suicidal) group (see Table III).

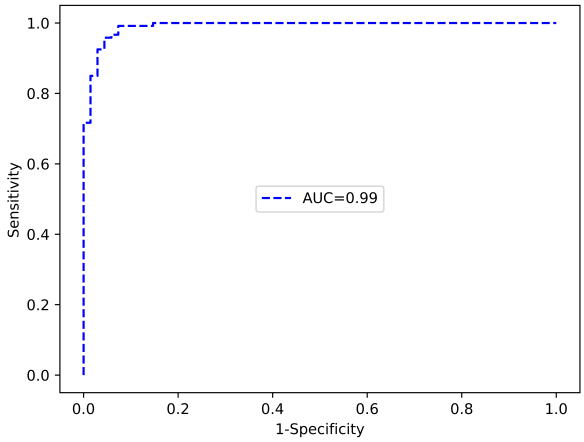

Fig. 1. Receiver operating characteristic (ROC) curve for the suicide risk score.

TABLE III
SUICIDAL RISK RESULT FOR THREE MODELS.

Model/data	Precision	Recall	Specificity	F1-score
Linguistic Indicator Approach Suicidal risk (38) Normal group (5419)	0.12	0.9	0.96	0.21
SVM Suicidal risk (38) Normal group (5419)	0.92	0.89	1.00	0.91
RoBERTa Suicidal risk (38) Normal group (5419)	1.00	0.97	1.00	0.99

B. Detecting Suicidal Risk in the Wild

Our different approaches were tested on the incel graveyard dataset (Reddit; *u/IncelGraveyard*) in order to identify how many of these posts could be classified as a high risk of suicide. The SVM model classified 56% of the incel graveyard posts as a "high risk of suicide" and the RoBERTa classified 70% as a "high risk of suicide". The linguistic indicator approach classified 88% of the incel graveyard posts as a "high risk of suicide". These findings show that the linguistic indicator approach had the best performance of all tested models.

V. EDWIN – A DIGITAL ASSISTANT FOR SUICIDE RISK ASSESSMENTS

Following our aim, we turned to implementing the best available model into a practicable tool: a digital assistant that can be used to determine suicide risk based on written communications. The idea is to provide researchers and professionals in relevant domains, e.g. psychologists, psychiatrists, clinicians, professional counselors, social workers, and mental health-trained responders the opportunity to conduct an instant suicide risk assessment.

To this end, we chose to implement the model that performed best in the wild - the linguistic indicator approach - and created a model where the outcome is a suicide risk score that can be evaluated in relation to the thresholds that were established above. As a starting point, our diverse

datasets were analyzed: the training data, the test data and the data from the incel graveyard (Reddit). As mentioned above, we (a) generated a score for each text and each of the 11 variables/dictionaries and subsequently, (b) using the threshold established for the training data for the *linguistic indicator approach* model, transformed these 11 variables to binary variables (0/1 for absent/present), and (c) averaged the binary variables for each text and multiplied the average by 100 to arrive at a suicide risk score between 0 and 100 for each text. We then used the risk scores together with the binary variable (suicide note/positive classes vs. normal text/negative classes) and conducted a ROC analysis. The results revealed that the area under the curve (AUC) was .98 (SE = .003, 95% CI [.977, .990]). The optimal threshold of this model was a risk score of 45.45. Using this threshold, we could correctly classify 91% of the suicide notes as a "high risk of suicide" and 96% of the normal (non-suicidal) texts as a "low risk of suicide". Thus, using a binary classification, we had 4% false positives and 9% false negatives.

While it is acknowledged that the ROC curve analysis is one of the least arbitrary ways to deal with the cutoff in binary tasks, we still wanted to reduce the number of false negatives. One way to do this is to use the Traffic Light Protocol - an intuitive approach that provides a departure from the binary yes/no classification. Therefore, we primarily aimed to keep false positive and false negative cases below 5%. A series of analyses showed that the Green class included 91% of the normal (non-suicidal) texts and only 4.5% of the suicide notes, while the Red class included 91% of the suicide notes and 4% of the normal (non-suicidal) texts. The Yellow class included 4.5% of the suicide notes and 4.8% of the normal (non-suicidal) texts (see Table IV).

TABLE IV
RISK SCORE RANGE FOR THE TRAFFIC LIGHT PROTOCOL CLASSES AND PROPORTION OF THE CASES WITHIN EACH CLASS IN THE DATA PRESENTED IN THIS STUDY.

Risk Score Range	Traffic Light Classification	% Classified	
		Suicide text	Non-Suicide text
0-29	Green	4.49	91.19
30-39	Yellow	4.49	4.82
40-100	Red	91.02	3.99

VI. Discussion and Limitations

The results reported above showed that the machine learning approaches performed better when classifying suicide letters as a "high risk of suicide" than the linguistic indicator approach. However, when testing our models in the wild, on a new unseen dataset, the linguistic indicator approach performed considerably better than both machine learning approaches. This suggests that identifying indicators/dictionaries of suicide is an approach that would need further investigation and attention. The process of theoretically identifying key indicators and incorporate these in a model proved successful.

Our suicide risk assessment assistant, *Edwin*, was built with the intention of preventing and intervening suicide by understanding and analyzing the suicidal mind through a language lens. The results show the potential of utilizing text analysis to assess the risk of lethal self-injury in written communications, and thus evaluating suicidal tendencies of their authors. Based on patterns in communications that express suicidal intentions, we were able to construct a general diagnostic tool that can be used in a variety of real-world environments. However, it needs to be stressed that the objective tool should not be used as a substitute for a full clinical assessment, and should only be interpreted as a preliminary instrument that can assist in determining if the written communication at hand contains alerting phrases that might indicate that its author is at risk of committing suicide. If the analysis of the tool results in a "high risk of suicide", a licensed suicidologist should conduct a full suicide risk assessment of the author. If the result is a "low risk of suicide", the author may nonetheless require a full suicide risk assessment, depending on the context of concern that is warranted by *Edwin* – such as a clear danger to others or an indication of a mental disorder.

In recent years, machine learning techniques have been applied to several domains related to health and social issues. Despite remarkable recent advances, it is important to note that many, if not most, of the resulting models are still lacking an understanding of the meaning of the data that they process. This is also valid for text analysis models. In many settings, these models cannot reach human-level accuracy, and problems may also occur when applying machine learning models on new unseen data. Differences between training data and real-world input can affect the performance in ways that are difficult to predict.

In addition, dictionary-based approaches have their limitations. First, the meaning of words can be context dependent, suggesting that words may have different meanings depending on how they are used. Also, dictionary-based analyses employ dictionaries that are defined a priori, without fully considering the domain that they are supposed to be used in. This means that the analysis may be sensitive to vocabulary variation that is introduced by, for example, slang words, different spellings and domain-specific terminology. Inability to handle vocabulary variation increases the risk of underestimating which in turn might lead to an inaccurate analysis.

Another challenge that complicates automatic suicide risk detection is that there are a very limited number of benchmarks for training and evaluating suicide risk detection.

Overcoming challenges such as discrepancy between training data and real-world, vocabulary variation across domains, and the limited number of benchmarks, requires extensive and methodical multidisciplinary research programs. The research we report here is hopefully one step in this direction.

VII. Acknowledgements

We thank Dr. John Olsson from the Forensic Linguistics Institute for providing 200 genuine suicide notes to our database for this project. The computations were enabled by resources provided by the Swedish National Infrastructure for Computing (SNIC) at Uppsala University, partially funded by

the Swedish Research Council through grant agreement no. 2018-05973.

REFERENCES

[1] G. Turecki and D. A. Brent, "Suicide and suicidal behaviour," *Lancet*, vol. 387, no. 10024, p. 1227–1239, 2016.

[2] D. Klonsky. (2021) 404 error: Mistakes we need to stop making in suicidology. [Online]. Available: https://cams-care.com/resources/events/404-error-mistakes-we-need-to-stop-making-in-suicidology-webinar

[3] S. Langer, J. Scourfield, and B. Fincham, "Documenting the quick and the dead: A study of suicide case files in a coroner's office," *The Sociological Review*, vol. 56, no. 2, pp. 293–308, 2008.

[4] S. Timmermans, "Suicide determination and the professional authority of medical examiners," *American Sociological Review*, vol. 70, no. 2, pp. 311–333, 2005.

[5] S. T. Black, "Comparing genuine and simulated suicide notes: a new perspective," *Journal of consulting and clinical psychology*, vol. 61, no. 4, p. 699–702, 1993.

[6] B. A. Messner and J. J. Buckrop, "Restoring order: Interpreting suicide through a burkean lens," *Communication Quarterly*, vol. 48, no. 1, pp. 1–18, 2000.

[7] E. S. Shneidman and N. L. Farberow, "Clues to suicide," *Public Health Reports (1896-1970)*, vol. 71, no. 2, pp. 109–114, 1956.

[8] E. S. Shneidman, *The suicidal mind*. New York: Oxford University Press, 1996.

[9] E. S. Shneidman, *Autopsy of a suicidal mind*. New York: Oxford University Press, 2004.

[10] J. Olsson and J. Luchjenbroers, *Forensic Linguistics*, 3rd ed. London: Bloomsbury Publishing, 2014.

[11] American Association of Suicidology, "Warning signs," https://suicidology.org/resources/warning-signs/, accessed: 2020-10-08.

[12] National Institute of Mental Health, "Warning signs of suicide," https://www.nimh.nih.gov/health/publications/warning-signs-of-suicide, accessed: 2020-10-08.

[13] S. Ji, S. Pan, X. Li, E. Cambria, G. Long, and Z. Huang, "Suicidal ideation detection: A review of machine learning methods and applications," *IEEE Transactions on Computational Social Systems*, vol. 8, no. 1, pp. 214–226, 2021.

[14] J. Lopez-Castroman, B. Moulahi, J. Azé, S. Bringay, J. Deninotti, S. Guillaume, and E. Baca-Garcia, "Mining social networks to improve suicide prevention: A scoping review," *Journal of Neuroscience Research*, vol. 98, no. 4, p. 616–625, 2020.

[15] Y. Ophir, R. Tikochinski, C. Asterhan, I. Sisso, and R. Reichart, "Deep neural networks detect suicide risk from textual facebook posts," *Scientific reports*, vol. 10, no. 1, 2020.

[16] R. Sawhney, P. Manchanda, R. Singh, and S. Aggarwal, "A computational approach to feature extraction for identification of suicidal ideation in tweets," in *Proceedings of ACL 2018, Student Research Workshop*. Association for Computational Linguistics, 2018, pp. 91–98.

[17] J. W. Pennebaker, M. E. Francis, and R. J. Booth, "Linguistic inquiry and word count (LIWC): A text analysis program," *Mahwah: Lawrence Erlbaum Associates*, 2001.

[18] Q. Cheng, C. L. Kwok, T. Zhu, L. Guan, and P. Yip, "Suicide communication on social media and its psychological mechanisms: An examination of chinese microblog users," *International journal of environmental research and public health*, vol. 12, pp. 11 506–27, 09 2015.

[19] J. Pestian, H. Nasrallah, P. Matykiewicz, A. Bennett, and A. Leenaars, "Suicide note classification using natural language processing: A content analysis," *Biomedical informatics insights*, vol. 2010, no. 3, pp. 19–28, 2010.

[20] J. Shapero, "The language of suicide notes," Ph.D. dissertation, University of Birmingham, 2011.

[21] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018.

[22] S. Daly and A. Laskovtsov, ""Goodbye, my friendcels": An analysis of incel suicide posts," *CrimRxiv*, 2021.

[23] A. Shrestha, N. Akrami, and L. Kaati, "Introducing digital-7: Threat assessment of individuals in digital environments," in *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2020, pp. 720–726.

[24] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," *CoRR*, vol. abs/1907.11692, 2019.