The Hitchhiker's Guide to Computational Linguistics in Suicide Prevention

Accepted Manuscript (May 2, 2021) – *Clinical Psychological Science*

Yaakov Ophir[1]

Refael Tikochinski[1]

Anat Brunstein Klomek[2]

Roi Reichart[1]

[1]Technion – Israel Institute of Technology, [2]Interdisciplinary Center Herzliya

Authors' Note

Correspondence should be addressed to Yaakov Ophir.

E-mail: yaakov.ophir@campus.technion.ac.il

Abstract

Suicide, a leading cause of death, is a complex and a hard-to-predict human tragedy. This article introduces a comprehensive outlook on the emerging movement to integrate Computational Linguistics (CL) in suicide prevention research and practice. Focusing mainly on the state-of-the-art Deep Neural Network models, this "travel guide" article describes, in a relatively plain language, how CL methodologies could facilitate early detection of suicide risk (section 1). Major potential contributions of CL methodologies (e.g., word embeddings, interpretational frameworks) for deepening our theoretical understanding of suicide behaviors (section 2) and promoting the personalized approach in psychological assessment (section 3), are presented as well. Importantly, the article also discusses principal ethical (section 4) and methodological (section 5) obstacles in CL-suicide prevention, such as the difficulty to maintain peoples' privacy/safety or interpret the "black box" of prediction algorithms. Ethical guidelines and practical methodological recommendations addressing these obstacles, are provided for future researchers and clinicians.

Keywords: Suicide Prevention; Computational Linguistics; Natural Language Processing; Deep Neural Networks; Machine Learning; Artificial Intelligence

Suicide, one of the most common and painful human tragedies, is a complex and somewhat enigmatic phenomenon (Hedegaard et al., 2018; Turecki et al., 2019). Although the current version of the Diagnostic and Statistical Manual of Mental Disorders (DSM-5), singles suicide behaviors as a possible distinct condition for future diagnostic manuals (American Psychiatric Association, 2013), suicide is still not considered a well-defined medical disease (Turecki et al., 2019). Theoretical psychologists struggle to reconcile this seemingly irrational behavior of self-destruction with the fundamental assumption of evolutionary theory (i.e., to survive and reproduce) (Aubin et al., 2013), and empirical researchers struggle to develop prediction models that will improve our ability to detect suicide risk and provide psychosocial help to individuals in need (Franklin et al., 2017). In fact, a wide-scope meta-analysis of five decades of suicide research revealed that our current ability to predict suicide ideation or behaviors is "only slightly better than chance" (AUCs = 0.56 – 0.58) (Franklin et al., 2017). Unfortunately, in reality, most individuals who suffer from suicide ideation are not identified by the mental health system and do not receive psychosocial care (Bruffaerts et al., 2011).

One of the reasons that suicide prediction is such a complicated task is that its etiology involves complex interactions between multiple bio-genetic, demographic, psychological, and social risk factors (Levi-Belz et al., 2019; Turecki et al., 2019). The full set of risk factors is usually not available to the researcher, and the existence of single risks, informative as they may be, provides only little predictive value. We know, for example, that most people who attempt suicide have a mental disorder, but since the reversed proposition is not true (i.e., most people with mental disorders do not attempt suicide), the relative contribution of this risk factor to suicide prediction is low (Franklin et al., 2017; Ribeiro et al., 2019). Moreover, the criterion measure of suicide risk may be hard to predict because suicide ideation and its associated risk factors can fluctuate dramatically over short periods of time (even within few minutes/hours) (Kleiman et al., 2017). This instability in suicide ideation, whereby actual suicide attempts may occur with very little warning or planning, contributes to the timely concerns that individuals at risk will engage in hasten life-threatening behaviors in response to the isolation and hardships that were triggered during the COVID-19 pandemic (Banerjee et al., 2020; Gunnell et al., 2020; Klomek, 2020).

To overcome these inherent obstacles in suicide prevention and break through the prediction ceiling in the current state of the literature, more and more scholars are recommending to integrate research methodologies from the field of Machine Learning (ML) – the study of computer algorithms that "learn" to preform prediction tasks based on training data that reflects past experience (Alpaydin, 2020). Within large datasets, ML algorithms are able to identify complex patterns, assess multiple risk factors simultaneously, and form accurate predictions regarding other, unseen data, in ways that were unattainable in the past. Notably, a specific field of research that makes use of ML strategies and that has been proven to be highly relevant to suicide prevention is Computational Linguistics (a.k.a. Natural Language Processing[1]) (Resnik et al., 2020) – a growing field that integrates knowledge and tools from multiple disciplines, such as linguistics, cognitive science, and artificial intelligence, to process and model human language (Clark et al., 2013).

In practice, Computational Linguistics (CL) tools allow researchers to analyze large amounts of texts and construct computerized algorithms that extract valuable, and in some cases hidden, patterns of meaning from human language. Not surprisingly, these new abilities have been harnessed by researchers to predict a range of public health issues based on textual data solely (Paul & Dredze,

[1] The term Computational Linguistics (CL) is often used as a synonym to Natural Language Processing (NLP). Although the jargon expression 'NLP' seems more common among computational scientists, it is often used in the specific context of technology-oriented research. This is in contrast to the term CL, which is informally considered as a more general and more inclusive expression that also contains language analyses from human-related perspectives.

2017). These include medical concerns such as infectious diseases, obesity, and substance use, as well as mental health conditions (Paul & Dredze, 2017; Perna et al., 2018; Shatte et al., 2019). Remarkably, considerable research has been directed at the detection of suicide ideation/behaviors (for updated reviews, see in: Bernert et al., 2020; Burke et al., 2019; Ji et al., 2020) and depression (Giuntini et al., 2020; Guntuku et al., 2017), one of the most dominant risk factors for suicide behaviors, and a major psychiatric and economic burden by itself (Clark & Beck, 2010; Evans-Lacko & Knapp, 2016).

The rationale behind this new research trend is that written or spoken human language contains explicit, as well subtle, signals about people's thoughts, feelings and behaviors (Fasold, 1990; Wardhaugh, 2011). Specific words and language patterns may tell us about the speaker's state of mind even when they are not overt. Depression, for example, can be tracked by explicit manifestations of negative emotions and depressive symptoms (e.g. sad, worthless, cry, lonely) as well as by more implicit language signals such as excessive usage of first person pronouns (e.g. I), past tense verbs (e.g. had), somatic complaints, and hostility (De Choudhury et al., 2013; Eichstaedt et al., 2018; Rude et al., 2004). Similarly, suicide risk can be identified using explicit suicide-related expressions (e.g., "kill myself", "wish I was dead") as well as by more subtle contents, such as negative emotions and physical complaints (Ji et al., 2020; Ophir, Tikochinski, et al., 2020).

As can be seen from some of the above citations, the idea that language can be leveraged for detection of mental health conditions is not entirely new. Since 1999, dozens of studies aimed to predict suicide risk using various ML and CL methodologies. According to a recent literature review, a total of 87 studies implemented ML investigations in suicide prevention research, of which 29 utilized CL techniques (Bernert et al., 2020). However, the vast majority of these CL studies have been published in the past six years. In fact, only four out of the 29 CL studies were published before

2015 (Bernert et al., 2020). Notably, this period of time, according to many computational scientists, marks the "revolution" of deep learning algorithms (Sejnowski, 2018) – computational models that rely on multiple layers (hence "deep") of artificial neural networks, also known as Deep Neural Networks (DNN) (Goodfellow et al., 2016). While these models have already been proposed in the 1950s, they have only become prominent in the last decade, with the increased accessibility to big data online (e.g., news websites, blogs, and social media) and the development of strong computers with Graphics Processing Units (GPUs) that could handle the heavily parameterized DNN models. In fact, the DNN revolution had a substantial impact on a variety of ML-related fields, including, for example, computer vision (Szegedy et al., 2016) and bioinformatics (Webb, 2018).

Specifically, a particular subgroup of DNN models that have become the state-of-the-art methodology in CL are the large, pre-trained, attention-based language encoders, such as BERT (Devlin et al., 2018) and GPT-3 (Brown et al., 2020) (for further description of this type of DNN models, see section 2.1). These powerful language encoders influenced both the scientific and the industrial communities, enhancing most CL applications (Rogers et al., 2020). Soon enough, these new CL tools have been utilized to predict a wide range of human behaviors (e.g., Ben-Porat et al., 2020; Oved et al., 2020). It is therefore expected that more and more researchers will adopt these promising CL tools, thus changing the way we conduct research in social and psychological sciences (Sejnowski, 2018).

A complementary historical change that boosted the use of CL technologies has been the emergence of the extremely popular social networking platforms (e.g., Facebook, Twitter, Instagram, and Reddit). While only 20 years ago writing was used mainly for formal communication purposes (e.g., research, literature, bureaucracy, and diplomacy), today, with the dawn of the Internet and the social media era, writing has become an integral part of informal, everyday activities (McCulloch, 2020). In the US for example, approximately 90% of teenagers and young adults, and 75% of older adults (30-64), use at least one social networking site (Pew Research Center, 2019). In practice, the various social media platforms, which contain large amounts of everyday written texts, have become an unprecedented source of psychological information about the person behind the keyboard (Kosinski et al., 2015; Mneimneh et al., 2021; Schwartz et al., 2013). Indeed, most of this information is not expected to include explicit references to suicide ideation or to mental health diagnoses, since many users refrain from disclosing sensitive information about themselves (Ophir, 2017; Ophir et al., 2019). However, researchers who employ CL methods may find hidden, non-explicit language patterns that will distinguish at-risk users from non-suicidal ones. Correspondingly, 15 studies, out of the 29 CL studies mentioned above, aimed to predict suicide from social media data (Bernert et al., 2020), thus extending our ability to detect suicide risk from everyday language, even among non-clinical populations.

Goals and structure of the current article

Altogether, the rise of the social media and the powerful CL methodologies opened unprecedented opportunities to conduct large-scope and complex investigations of human behaviors. The goal of this article is to introduce to mental health scientists, clinicians, and policy decision makers, who may be less familiar with current computational technologies, the fundamental concepts and the principal benefits and limitations of CL research and practice in the field of suicide prevention. Importantly, this article does not downplay the inherent obstacles of CL methodologies. On the contrary, it aims to lift their somewhat shiny cover and provide a wide outlook on their intricate mechanism, while maintaining a relatively simple terminology. Building upon prior reviews and perspective articles (Bernert et al., 2020; Burke et al., 2019; Ji et al., 2020; Resnik et al., 2020), the current overview presents a comprehensive discussion of the major potential benefits and limitations of key CL methodologies in the context of suicide prevention. While this overview article cannot substitute a technical manual, it can be used as a relatively easy to follow, "hitchhiker's guide"[2] to the fascinating galaxy of CL usages in suicide prevention. To obtain an integrative and complete picture of the field, interested readers are encouraged to read this travel guide chronologically, as a whole, since each section builds upon its predecessors. However, readers can also navigate between specific segments of interest, since every section in this article may stand on its own, and inter-references to other sections with relevant complementary information are provided throughout the article.

The structure of the article is as follows: Section 1 offers an integrative review of the existing literature on CL in suicide prevention and describes the primary potential of CL methodologies for early detection of suicide risk. This section presents key DNN models that suit this cause, outlines the unique characteristics of these models that are relevant to suicide detection, and suggests three ways in which CL could improve our practical ability to detect suicide risk. The remaining four sections of the article provide a substantial extension of the existing literature reviews and perspective articles. Specifically, section 2 provides an overview of how CL could advance our theoretical understanding of suicide ideation and behaviors – a subject that is less discussed in the literature. This section describes three principal CL methodologies, which could lead to scientific discoveries in the field of suicide prevention: (a) the state-of-the-art geometric language representations (i.e., static and contextualized DNN-based word embedding models), (b) the well-established and highly used non-DNN semantic analysis of topic modeling, and (c) the emerging interpretational frameworks for

[2] The title of this article is inspired by Douglas Adams' science fiction series "the hitchhiker's guide to the galaxy", in which the "guide" refers to a fantastic electronic book that provides easy to follow (and humoristic) background information regarding complicated concepts in the universe.

DNN research. Then, section 3 illustrates how the CL methodologies, which were described in the previous sections, can promote a personalized approach, both in suicide prevention and in the more general field of psychological assessment. Specifically, this section discusses three ways in which CL could be utilized for addressing principal issues in psychological assessment, such as the wide heterogeneity that exists within mental health diagnoses, the reliability of subjective self-report information, and the difficulty to conduct indirect, yet psychometrically validated assessments.

Following the presentation of the potential benefits, we move to discuss the ethical and methodological limitations of CL applications in suicide prevention. Section 4 presents the complex ethical challenges that are bound to arise in CL-based suicide prevention, which involves both sensitive content (suicide risk) and controversial procedures (automated monitoring). This section offers ethical guidelines to maintain the privacy, autonomy, and safety of individuals at risk, both in research and in real-life settings, and discusses other ethical issues that are relevant specifically to CL technologies (e.g., implicit mental health stereotypes in language processing models). Finally, section 5 outlines key scientific challenges in CL-based suicide prevention research and practice, such as the heterogeneity in suicide ideation and behaviors, the domain adaptation problem, and the difficulty to interpret the "black box" of DNN-based prediction models. In this section, we propose practical solutions to these issues, which could increase the generalizability and interpretability of CL prediction models, and provide methodological recommendations for future research and practice.

1. Facilitating early detection of individuals at risk for suicide

The first and primary contribution of the DNN revolution to CL suicide research is a substantial improvement in text classification task performances (Schafer et al., in press), an improvement that may increase our practical ability to detect individuals who suffer from suicide ideation. Key DNN models, which are suitable to this cause are: (1) Long Short Term Memory

(LSTM) networks (Hochreiter & Schmidhuber, 1997) and their bidirectional counterpart (Bi-LSTM) models, that account for the sequential nature of texts (Zhou et al., 2016); (2) deep convolutional networks (Conneau et al., 2017), which exploit the linguistic signals encoded in the physical proximity of words, sentences and paragraphs in the text; and (3) the advanced attention-based transformers (Vaswani et al., 2017), that can model dependencies between words, phrases, and sentences in various parts of the text. In this section, we describe principal ways in which these methods and similar ones could contribute to early detection of suicide risk.

1.1. Improving suicide prediction performances. The first notable strength of CL methods in suicide research concerns their high prediction performances compared with traditional statistical approaches (Schafer et al., in press). Whereas top-down, traditional studies (that do not use CL or other ML methods) usually investigate a small number of pre-defined, theory-driven independent variables, bottom-up CL models usually extract multiple language patterns from large datasets, and integrate the wide-ranging information encoded in these patterns into their prediction mechanism. Other, non-textual data such as images and emoji, may be integrated as valuable inputs to these models as well, which could improve the final predictions. Moreover, these models are particularly effective for making joint predictions from multiple input variables (e.g., Oved et al., 2020) on multiple dependent variables (Ruder, 2017). This joint modeling of multiple output variables is facilitated in DNN-based algorithms because they can constitute a joint objective function (with components related to each of the output variables), which is optimized in a unified continuous optimization framework. This quality is especially relevant for the complex task of suicide prediction, which could involve multiple types of textual and non-textual predictors from various sources (e.g., personal texts, clinical interviews, psycho-diagnostic information, and behavioral observations) alongside multiple output variables, including a range of suicide ideation and behaviors as well as additional psychological and psychiatric risk factors, such as rumination or depression (for a recent example of a multi-task CL model that jointly predicts a hierarchical set of risk factors, see in Ophir, Tikochinski, et al., 2020).

Without underestimating the limitations of CL models (see section 4 and section 5), it seems that this distinct quality of DNN-based CL models facilitated the emergence of a series of studies that resulted in promising prediction performances of suicide risk (e.g., Coppersmith et al., 2018; Ophir, Tikochinski, et al., 2020; Roy et al., 2020; Zheng et al., 2020; Zirikly et al., 2019) as well as of depression (e.g., De Choudhury et al., 2013; Eichstaedt et al., 2018; Guntuku et al., 2017; Tadesse et al., 2019), which, in many cases, precedes the actual suicidal behavior (American Psychiatric Association, 2013; Hawton & van Heeringen, 2009). In fact, recent reviews of studies that integrated CL methods in suicide research showed that these methods accomplish high quality predictions ($.61 \leq AUC \leq .95$) (Bernert et al., 2020; Burke et al., 2019), which significantly outperform the existing, close-to-chance-level predictions made by more traditional statistical methods (Schafer et al., in press). Thus, although this field of research is still in its infancy, the current state of the literature suggests that CL algorithms should be considered as accurate and powerful practical tools for suicide risk detection.

1.2. Expanding suicide screening to large and detached populations. A second advantage of CL methods concerns their automatability and potential diffusion to large populations. Whereas traditional suicide screening tactics (e.g., self-report questionnaires), which are being applied in primary care services, hospitals, and schools (Horowitz & Ballard, 2009; Thom et al., 2020), are expensive and limited in their ability to reach detached populations, CL-based technology allows the creation of automatic monitoring tools, which could be applied in multiple, everyday environments. Opportunely, the emergence of the two contemporary technologies discussed in the introduction (i.e., the ubiquitous social media and the deep learning language processing algorithms) has been leveraged by many researchers to develop ML and CL suicide prediction models that are based on the numerous non-medical, everyday language that is accumulating online (Burke et al., 2019).

Still, this does not imply that identifying suicide risk from everyday language on social media is a straightforward task (Olteanu et al., 2019). As opposed to formal medical records or other clinical data (e.g., responses to psychological questionnaires), everyday language on social media is extremely "noisy". Different network sites (e.g., Facebook vs. Instagram) may be characterized with dissimilar contents (e.g., texts vs. pictures) and users' demographics (e.g., adolescents vs. older adults) (Pew Research Center, 2019); and users' activity may include a great amount of irrelevant information, with very little explicit statements regarding their mental state. In fact, many social media users do not feel comfortable discussing their mental health diagnosis in public and refrain from sharing their depressive or suicidal ideation in an explicit manner (Ophir, 2017; Ophir et al., 2019). Therefore, simple algorithms that search for specific words or expressions (e.g., "I wish I was dead"), may not be appropriate for this task.

Fortunately, the multi-task nature of DNN-based CL tools and their ability to extract hidden, non-explicit textual and non-textual signals from input data (see section 1.1 and section 2.1), could be leveraged by researchers to overcome this obstacle and distinguish at-risk users from non-suicidal users, based on more implicit information. However, in order to extract this hidden information, CL researchers are required to collect offline external validations of suicide risk, such as medical records or other clinically valid measures, which can serve as the 'ground truth' criterion of suicide risk in the learning phase of the algorithm (Chancellor & De Choudhury, 2020). In other words, since they are not searching for (rare) explicit references to suicide, they cannot rely solely on ground truth criteria which consist of explicit online content that is judged (by experts or by non-experts) to be indicative of suicide risk (see also section 5.3).

In our recent work, for example, we established our ground truth labels of suicide risk through a rigid data collection procedure, in which social media users had to complete validated psychological questionnaires including the Columbia–suicide severity rating scale (Posner et al., 2011). We then trained deep neural network models to predict these labels from the users' (independent) social media texts (Ophir, Tikochinski, et al., 2020). Interestingly, the results of this process indicated that CL prediction models do not necessarily require explicit suicidal language to form accurate predictions. Instead, they can rely on subtle language differences between suicidal and non-suicidal users, such as emotionally charged wordings and distinct topics (e.g., references to negative emotions and experiences among suicidal users and references to religious/spiritual experiences among non-suicidal users) (Ophir, Tikochinski, et al., 2020). A similar conclusion was formulated by Coppersmith et al. (2018) who argued that CL methods "depend on a wide swath of subtle clues, rather than a few indicative phrases" (p. 4). This sophisticated ability to extract abundant distinctive language patterns that are less visible to the human eye (see also section 2.1), alongside the wide expansion of written language usage online, can significantly improve our ability to detect suicide risk even among detached populations that do not receive regular psychosocial care.

1.3. Providing real-time indications of suicide risk. Finally, as mentioned in the introduction, one of the reasons that suicide prediction has been such a complicated challenge is that acute and imminent suicide risk is not a constant psychological state. Suicide ideation ranges from general and somewhat amorphic thoughts about death, to an ambivalent position whether life is worth living, through thoughts about suicide, and finally to a decisive resolution (even if impulsive) to kill oneself (Galynker et al., 2017; Posner et al., 2011). For example, a person at a relatively low suicide risk, who tends to engage in recurring depressive thoughts about the futility of life, may experience a sudden intense emotional pain after a traumatic event, which will put him in an imminent risk of suicide. Notably, CL tools can be utilized for a continuous monitoring of suicide risk in real time. For example, a study that analyzed the language used by members of an online support community showed that computational models can identify shifts from a general mental health discourse to a more concrete suicide related discourse (De Choudhury et al., 2016).

Although CL algorithms that do not include temporal information are also capable of producing high quality predictions of suicide risk, the integration of such temporal information may provide us a more nuanced perspective on suicide behaviors (Burke et al., 2019), and eventually improve our ability to detect and prevent them on time. For instance, Roy et al. (2020) who used neural networks to predict the appearance of explicit warning signs on Twitter (e.g., a tweet, in which the user writes: "I am planning to kill myself") found that the quality of the prediction improved as the data used for the prediction (i.e., the previous textual activity of the user) was closer in time to the warning sign (Roy et al., 2020). Indeed, high quality predictions were also observed in this study regardless of the investigated timeframe (i.e., the proximity in time of the Twitter activity to the specific suicidal tweet). However, the best prediction was obtained when the starting position of the analysis was one day prior to the explicit warning sign. Importantly, the above described DNN models that rely on LSTM and Bi-LSTM networks are especially relevant for this task because they account for the sequential nature of the input. This quality allows these models to capture crucial information regarding the chronology of the users' online activity and the actual time in which they posted warning signs or other relevant signals that contribute to the final prediction of the suicide risk.

In conclusion, not only that CL tools can improve the accuracy of suicide prediction models (section 1.1) and increase our accessibility to individuals at risk who lack psychosocial support (section 1.2), but they can also contribute to our efforts to monitor the risk in real time (section 1.3). Future application of such tools among large populations may therefore expand early suicide detection efforts in the community, encourage at-risk individuals to seek help, and hopefully contribute to a significant reduction in suicide rates around the world.

2. Deepening our theoretical understanding of suicide ideation and behaviors

Notably, the potential benefits of CL for suicide prevention extend far beyond the early detection described above. A second major potential benefit of CL concerns our basic perception of suicide ideation and behaviors. In this section, we describe *three* principal CL methodologies (i.e., geometric language representations, topic models, and interpretational techniques), which could lead to scientific discoveries and deepen our theoretical understanding of suicide and its related risk factors. A better understating of suicide, we believe, could eventually improve our ability to provide mental health patients with more accurate and more nuanced diagnoses, and therefore with a more suitable psychosocial care (see also section 3).

2.1. Geometric language representations. One of the greatest achievements of CL research in the last decade is the development of effective representation models for words (a.k.a. word embeddings), as well as for larger texts. Contemporary DNN-based word embedding models produce vectors, consisting of a set of coordinates (entries) that represent the location of the word in the semantic space. This semantic space is built by the model such that semantically similar words (e.g., the words 'cat' and 'dog') should be located in a close proximity to each another. Indeed, the foundation of this CL methodology is not new. Computational linguistics have long been trying to represent language data through property (feature) vectors. However, in the past, these vectors were typically extremely large and sparse (i.e. they consisted of numerous entries, many of which with the value of 0). For example, a traditional word embedding strategy is to represent every word with a vector in which every coordinate corresponds to the co-location of this word with any other word in the lexicon, within a given textual window (e.g. in the same sentence, or in a 3-word window), where the co-location counts are based on large textual corpora. While such word representations may effectively represent the semantics of words (in the sense that semantically similar words are located in a close proximity in the semantic space), a lexicon of 5000 words for example, would yield under this method a vector representations of 5000 features, of which most would receive the value of 0, as most of the words in this lexicon do not co-locate with each other in large textual corpora. If we move to represent each word by its co-location with word pairs, the number of coordinates would increase to 25,000 and the number of 0 entries is likely to be even larger. Such vectors are very hard to interpret and their computer memory demands are very high. In contrast, contemporary DNN-based word embeddings are of a much lower dimension (up to several hundred coordinates compared to the tens of thousands coordinates of the previous representation methodology) and are much more dense (i.e. they contain only a small number of 0 entries). This compact representation makes computation and memory demands with these vectors much more modest. Importantly, they also provide a better semantic representation for words and eventually also for larger texts.

While the DNN-based word embedding literature is relatively young, as the first prominent model was introduced in 2013, computational linguists identify two developmental waves of this methodology. The first wave relates to the development of *static word embedding models*, such as 'word2vec' (Mikolov et al., 2013) and 'Glove' (Pennington et al., 2014), which are trained on large text collections (billions of words) and generate a single embedding vector for each word/phrase, regardless of its context. Indeed, this strategy does not distinguish between different senses of same words, such as in the word 'pass', which in some contexts means 'move' and in others means 'die'. However it is still an effective tool for generating valuable input features for supervised classifiers[3] (section 1) and for extracting word and phrase meaning from texts through their semantic spaces (Bolukbasi et al., 2016; Garg et al., 2018).

The second wave of DNN-based geometrical representation strategies introduced models that assign a distinct vector representation for words/phrases, depending on their textual context. These *contextualized word embedding models* are based on the aforementioned architectures (see the introduction and section 1.1), which include the Bi-LSTM (Peters et al., 2018), and most importantly, the flexible attention architectures of BERT (Devlin et al., 2018), GPT-3 (Brown et al., 2020), and T5 (Raffel et al., 2020), which are considered powerful tools that could model long-distance dependencies in texts. These novel contextualized models typically consist of millions to billions of parameters and their development and training process involves extremely large corpora (e.g. the entire Wikipedia). Briefly, the training procedure of these models is conducted through language-modeling objectives, whereby randomly selected words are deleted from the input text and the models' training goal is to predict the missing words based on their context. After training, these models can then be applied to new, unseen texts and generate appropriate contextualized vector representations for each word instance in these texts.

Importantly, the DNN models described above, and especially the contextualized embedding ones, may provide us an opportunity to deepen our theoretical understanding of complex human behaviors, such as suicide. Recent works showed that embedding models capture implicit knowledge

[3] Supervised classifiers refer to algorithms that are trained using labeled examples (e.g., a post that is labeled as a reference to suicide risk). These classifiers aim to learn the mapping function between the input and the desired outputs (e.g., the actual suicide risk). In contrast, unsupervised classifiers do not have labeled outputs. Their goal is therefore to extract features and patterns in a given (unlabeled) dataset.

regarding the semantic background of given texts, thus revealing hidden personal information about the individuals who wrote them. For example, a recent study used embedding models to map (cluster) textual data to distinct domains, thus allowing the researchers to determine the characteristics of the source of the data (e.g., medically-oriented texts) (Aharoni & Goldberg, 2020). Another notable example is a large inquiry that applied embedding models on 100 years of text data and found that the models captured gender and ethnic stereotypes (Garg et al., 2018). Correspondingly, embedding models that were trained on allegedly objective news articles also learned implicit gender stereotypes (Bolukbasi et al., 2016) (for a further discussion of the ethical aspects of these social biases, see section 4.3). This means that the impact of embedding models goes beyond plane predictions of certain phenomena, as they can be used to extract insights regarding the persons behind the texts. Suicide researchers can therefore utilize embedding models to explore the semantic space that distinguishes suicidal from non-suicidal individuals and offer insights regarding the unique cognitive and emotional characteristics of the at-risk group. For example, abstract notions, such as failure and sadness may appear in a closer proximity in the semantic (vector) space of suicidal individuals compared to non-suicidal individuals, thus perhaps implying of a cognitive distortion that is typical to depressive disorders (Beck, 1991). Other, more complicated semantic relationships may reveal new insights that could not have been hypothesized a-priori by the researchers, thus opening empirical windows to new theoretical perceptions regarding depression and suicide.

2.2. Topic Models. Another prominent, even if not very recent, CL methodology that could shed light on the unique semantic contents that occupy at-risk individuals is the Latent Dirichlet Allocation (LDA) Topic Modeling framework (Blei et al., 2003). Indeed, the 59-year-old field of CL (counting from the first annual meeting of the Association for Computational Linguistics) generated a large number of language processing methodologies, but given our focus on DNN methodologies and the limited scope of this overview, we chose to present only one non-DNN method. Notably, this last methodology of topic models is a very popular strategy to extract valuable and interpretable meaning from large collections of documents. Branching from the Bayesian statistical modeling framework, which has complementary advantages and disadvantages compared to the more recent DNN methodologies, topic models were applied in thousands of studies. In fact, the above citation by Blei et al. who introduced the LDA topic modeling method in 2003 has been cited over 36,000 times, to date.

In this unsupervised method, topics are defined as distributions over the words in the lexicon, such that these distributions are automatically learned from the data. The model represents each document as a mixture of topics – a vector of probabilities that quantify how strongly each one of the learned topics is associated with the document. In this method, the semantics of entire documents can be interpreted based on their unique associations with the extracted topics and the topics themselves can be interpreted or labeled with a title, based on their most associated words and phrases. A recent work on depression, for example, showed that topic modeling of Facebook texts can be used both for generating significant predictors of depression and for offering a semantic interpretation of these predictors (Eichstaedt et al., 2018). Specifically, in this study, topics that were strongly associated with words addressing sadness, loneliness, hostility, and somatic/medical complaints, were more prevalent among Facebook users diagnosed with depression, thus illuminating the emotional and cognitive preoccupations of this at-risk group. Complementing with this approach, analyzing the contents of automatically induced topics can reveal equally important protective factors that may help distressed individuals cope with their emotional difficulties. For example, in our previous work, we noticed (using a more basic interpretational tool than topic models) that religiously related contents were more distinctive of the non-suicidal group (Ophir, Tikochinski, et al., 2020) thus raising the hypothesis that a sense of belonging to a community and/or of a meaning in life plays a significant role in people's mental health maintenance. Notably, this insight could be integrated in actual therapeutic interventions that target suicide ideation.

2.3. Interpretational techniques. Finally, emerging interpretational frameworks for CL research may help researchers decipher the "black box" of the complex DNN-based prediction models (see section 5.1). The goals of these frameworks are to single out the most dominant features that allowed the computational model to make its' predictions and to assess to what extent the impact of these features can be generalized to other settings and datasets. Following are three representative interpretational techniques that can be utilized to achieve these tasks (for additional interpretational methods see in: Belinkov, 2021).

One way to extract potential explanations for DNN-based CL models is to examine the predictive power of different features. Like human perception judgements that are formed based on selective attention to different stimuli in the environment based on their perceived significance (e.g., a stimulus of a human face may have more value to the observer than the landscape behind it), attention-based text-representation tools, such as BERT, learn different predictive 'weights' to different features. Researchers can therefore utilize these different weights and offer explanations regarding the specific features that drove the model in its decision-making process (Jain & Wallace, 2019; Wiegreffe & Pinter, 2019). Another interpretational technique is to train a simpler model so that it could mimic the predictions of a larger and more complex model on input examples of interest. Indeed, these small models may not achieve the same prediction performance as the more complicated model, however, they can be significantly more interpretable because of their limited number of parameters (Sanh et al., 2019).

Finally, a third, and if we may say so, a highly promising interpretational method, which was developed by our research team to overcome inherent obstacles in CL-based interpretational research, such as shallow and misleading correlations (see also section 5.1) is the Causal Model Explanation through Counterfactual Language Models (CausaLM) (Feder et al., in press). CausaLM allows researchers to explore counterfactual language representation models and offer causal explanations even when the data have not been collected in a randomized controlled study. Briefly, this method relies on a careful subtraction of distinct information from the internal text representation of the model, and a further estimation of the pure effect of the remaining information. Typically, a causal graph (Pearl, 2009) is first built to outline all potential predictors and their internal interactions, and then one of these predictors is removed from the model, without impairing the remaining information (even though the remaining information may be intertwined with the removed variable). For example, a prediction model of suicide may assign a relatively large prediction weight to the gender of the author of a given textual input, simply because gender and depression (a close predictor of suicide) are highly correlated (American Psychiatric Association, 2013). This of course does not necessarily mean that gender *causes* suicide ideation. CausaLM would allow the researcher to subtract any textual information about the gender of the author from the internal text representation of the model while keeping all the information about the author's depression, and explore whether the prediction of the model regarding the suicide risk of the author changes. Although any findings that are extracted through the above described interpretational methods would require further investigations, preferably in rigid experimental designs, it is our belief that these methods could provide a "window to the person's soul" and uncover distinct psychological concepts that could shed light on the phenomenon of suicide.

3. Promoting a personalized approach in psychological assessment and suicide prevention

Importantly, the potential impacts of CL (sections 1 and 2) could extend far beyond suicide prevention. As described in the introduction, one of the reasons that led suicide researchers to turn to ML and CL methodologies was the complex and multicausal etiology of suicide. Knowing, for example, that a person is diagnosed with a mental disorder such as depression was not enough for making accurate suicide predictions because most diagnosed individuals do not attempt suicide. However, a complex etiology is not a unique characteristic of suicide behaviors. Since the appearance of the biopsychosocial model (Engel, 1977), the scientific consensus supports the position that mental illnesses, in general, may result from complex interactions between a wide range of biological, psychological, interpersonal, and environmental risk factors (Lehman et al., 2017). As stated in the DSM, "in the absence of clear biological markers or clinically useful measurements of severity for many mental disorders" (American Psychiatric Association, 2013, p. 21), an accurate psychological assessment is an extremely difficult task, both for humans and for computers (Hitchcock et al., 2021). Not only that different causes (e.g., a specific trauma versus a genetic predisposition or general poverty) can lead to similar disorders, in many cases, the manifestations of a given disorder can be quite heterogenous. Two individuals who suffer from major depression, for example, a well-recognized risk factor for suicide, may share very little, if any, joint symptoms, and yet be eligible to the same diagnosis, according to the DSM (Fried, 2017; Fried & Nesse, 2015).

To begin resolving this problem, contemporary psychological scientists seem to gradually adopt the emerging perception of personalized psychopathology – the art of tailoring the diagnosis and the intervention to the unique characteristics of the patient (Wright & Woods, 2020). This is in contrast to the more traditional perception, in which the patient is fitted into a "one size fits all/most" diagnosis and treatment (Woods et al., 2020). Indeed, the personalized approach is a growing trend in medicine in general (Collins & Varmus, 2015; Jameson & Longo, 2015), however, it is especially relevant to psychopathology research. This is because, in contrast to pure physiological conditions that might have distinct etiologies and symptomologies, the origins and manifestations of mental health conditions, as mentioned above, differ from person to person (Engel, 1981; Hyman, 2010). Moreover, even a small mental adversity (e.g., experiencing undesirable thoughts about failure) may have large and diverse effects on the individual's cognition, emotions, and daily behaviors. Opportunely, the need for a more personalized approach in clinical psychological science could be addressed today, with the rise of the DNN revolution and the availability of big data online (see the introduction). We propose that further research that will utilize the CL methodologies described above (sections 1-2) to develop nuanced and precise representations of idiosyncratic mental states (including suicide ideation) of individuals, could boost the personalized approach in psychological assessment in general and suicide prevention in particular.

Using CL methodologies to promote the personalized approach in suicide prevention can be manifested in a shift from 'main effect' studies (i.e., studies that aim to expose one or two major influences) towards a more complex view of suicide behaviors, which includes multiple risk factors, interactions between these factors, and diverged symptoms. Indeed, this complex view of suicide may be harder to interpret, however, the application of interpretational frameworks, topic models, or word embeddings described above, may prove to be clinically useful. Interpretational frameworks, for example, may help researchers understand why some depressed patients are more susceptible to suicide behaviors than others (e.g., depressed individuals with little familial and social support). Clinicians could then target their interventions at the specific facilitating/moderating psychosocial factors that were located by the CL researchers, to help their patients cope with suicide ideation.

In practice, CL-suicide researchers can aim to identify central mediating and moderating risk factors for suicide behaviors as well as meaningful interactions between these factors (Burke et al., 2019). They can also utilize the data-driven approach to identify new gross, yet psychometrically valid, clusters of human cognitions and behaviors as well as different subgroups of vulnerable populations. Potentially, these data-driven clusters may help researchers identify unique characteristics of specific populations at risk (e.g., depressed individuals who also suffer from a substance use disorder) and vulnerable age groups (e.g., adolescents, elderly). Moreover, if applied on sufficient and diverse data, CL methodologies may help researchers detect new (unknown) subgroups at risk, distinguish between different types and severity of self-harm behaviors, and differentiate between people who are responsive to usual psychological care and those who are less responsive. Altogether, the creation of various profiles of individuals at risk could help clinicians adjust their treatments to the specific needs and characteristics of their patients.

Using CL methodologies to promote the personalized approach in the more general field of psychological assessment can be manifested in the development of computerized tools for indirect, yet, psychometrically valid, psychological assessments. Today, most mental health diagnoses (e.g., depression) are given to patients', mainly based on their subjective reports of symptoms (e.g., depressed mood nearly every day in the past two weeks). Even when clinicians conduct structural or semi-structural interviews, and even when patients complete reliable and valid mental health questionnaires, the final diagnosis is still usually based on patients' reports, which are subjected to inaccurate perceptions, self-presentation biases, and, in some cases, even intentional distortions (e.g., patients who try to hide their suicidal ideation) (Paulhus & Vazire, 2007). To overcome these biases and to broaden the somewhat narrowed symptom-focused view of mental disorders, psychological scientists have been trying for decades to assess hidden psychological constructs and personality patterns through indirect projective psycho-diagnostic tools, such as the Rorschach test or the Thematic Apperception Test (TAT) (Groth-Marnat, 2009). However, these tools are often being criticized as having poor inter-rater reliability and poor criterion validity, and their usage in clinical/research settings is controversial (Groth-Marnat, 2009; Lilienfeld et al., 2001; Smith et al., 2018). Moreover, such indirect tools are applied only sparingly, among a small portion of the population because they require large investments of time and money.

Although somewhat futuristic, we expect that CL methods, which will be applied to large amounts of personal texts, would provide researchers and clinicians with a unique opportunity to conduct indirect psychological assessments without compromising essential psychometric assumptions. Potential texts for indirect psychological assessment may range from designated essays on specific themes written by patients at the request of their (CL-oriented) therapist, to personal diaries, online blogs, and, as mentioned above, social media texts. Notably, the last three types of texts are not generated in response to predefined psycho-diagnostic stimuli (e.g., a Rorschach stain) or to direct psychological questioning (e.g., a close-ended self-report scale), therefore increasing their authenticity, and perhaps also their psycho-diagnostic value. The informal language people generate online, for example, is not intended to please therapists or to answer direct questionnaires, and there is a reasonable possibility that it would give us a glimpse into the authentic psychological state of the user (Ophir, Rosenberg, et al., 2020). Researchers could therefore leverage this information to develop CL-based indirect assessment tools that could expose a rich and nuanced psychosocial profile of users, which will include for example delicate information regarding their emotional resources, cognitive biases, personality traits, social relationships, and internal parental representations (Groth-Marnat, 2009).

As mentioned above, DNN-based CL models are particularly effective for making joint predictions from multiple input variables on multiple dependent variables (section 1.1). They can be utilized to analyze posts, pictures, comments, and 'likes' and form predictions regarding a range of psychological features of the user. In this way, instead of using one or two unidimensional labels for diagnosing patients, computational psychologists could develop alternative classification methods and construct complex and idiosyncratic psychosocial profiles for each patient. These "round characters" may better reflect the contemporary biopsychosocial conceptualization of mental health conditions (Lehman et al., 2017) and provide clinicians with an informative list of traits, stressors, and resources regarding their patients. As long as clinicians are careful not to embrace these diagnostic products as absolute truths, they could utilize them for tailoring the therapeutic discourse to the unique psychological profiles of their patients. Ultimately, these potential CL research directions could broaden our perception of mental health conditions such as depression and suicide and contribute to the realization of personalized psychopathology.

4. Ethical considerations and principal guidelines

Alongside promising opportunities, technological advancements also bring complex ethical challenges. This general statement is especially relevant to CL-based suicide prevention, which involves both sensitive contents (suicide) and controversial procedures (automated monitoring). In fact, overreliance on big data and AI technologies is a prominent concern in the scientific and the public discourse. Some scholars and opinion makers are even warning that a continuous adoption of these technologies in our daily lives is bound to increase inequality and harm disadvantaged groups (O'Neil, 2016). Of course, a complete overview of all AI-related dangers is beyond the scope of the current article. However, in this section, we wish to discuss the principal ethical issues that are most relevant to research and practice in CL-based suicide prevention. Specifically, this section addresses four ethical dilemmas: (1) privacy considerations vs. the need for high quality prediction models, (2) the autonomy vs. the safety of suicidal individuals, (3) the benefits of real-life monitoring applications vs. the risks for intentional data abuse., and (4) the wish to utilize advanced CL systems vs. our limited knowledge regarding their internal mechanism.(e.g., the fact that CL models capture social stereotypes). In this discussion, we also propose several ethical guidelines for future researchers and developers, although we acknowledge that thorough scientific and public discussions are still very much required in this sensitive field.

4.1. Privacy considerations vs. the need for high quality prediction models. The first, and probably the most discussed, ethical concern in AI in general, and CL in particular, is the potential violation of people's privacy (Jobin et al., 2019). As mentioned in the introduction, a large portion of the studies that used CL methods to detect mental health conditions relied on publicly available data from social media, such as Twitter (Guntuku et al., 2017), that have relatively light privacy constrains. However, simple ethical solutions (i.e., justifying the wavery of privacy concerns by relying on publicly available data) may not be appropriate here because many social media users do not intend that their personal information will be used for research and do not assume that it will be seen by people outside their close social circles (Paul & Dredze, 2017). Moreover, even when studies are conducted in more secured networks (e.g., Facebook) private information may still leak out to unwanted parties and privacy violations may still occur.

In the infamous Cambridge Analytica scandal for example, private information from Facebook was originally collected for research purposes by academic personnel, but was then leaked, and abused by a third party, for non-scientific purposes (Isaak & Hanna, 2018). In another example of an ethically-controversial practice, Facebook users' news feeds were manipulated without their knowledge, to evaluate how positive/negative content affect users' well-being (Kramer et al., 2014).

Thus, in order to maintain participants' right to privacy, researchers are advised to adhere to consensual ethical norms, obtain informed consents from potential participants, and allow them to opt out whenever they choose to (Resnik et al., 2020; Verma, 2014). Alternatively, researchers may collect data from designated sources in which users provide an a-priori permission to collect and analyze their data, such as in the recently developed OurDataHelps.org infrastructure (Coppersmith et al., 2018). Although these designated sources come at the expense of investigating real and natural environments, they may encourage researchers to make progress in a relatively protected setting.

The privacy issue in CL-based suicide research, however, extends beyond the specific settings of a given study. A standard scientific norm in CL research (and ML research, in general) is to share the dataset that was used for developing the prediction model with the scientific community. This data sharing practice is essential to ensure the generalizability of the models. Indeed, the evaluation of models' generalizability is inherent to the basic research design of CL, which seeks to build modles that generalize from a train set and a test set. However, in this setup, both sets are typically derived from the same distribution. Further investigations of the models in various settings and datasets is therefore crucial (see also section 5.2). Alas, in suicide research, data sharing is highly problematic from an ethical point of view (Resnik et al., 2020). The collected psychosocial information about the participants is extremely sensitive and, in many cases, requires the researcher to assure the participants (at the data collection phase) that their information will not be transferred to others, even if these 'others' belong to a recognized scientific institution.

Consequently, CL models developed for sensitive tasks, such as suicide detection, are often evaluated on a (very) limited number of datasets, thus leaving the field of computational clinical psychology behind the state of the art in other CL research areas (Resnik et al., 2020). This posits a heavy burden on the capability of the research community to develop wide-coverage models (section

5.2). To overcome this obstacle, an emerging ethical approach suggests that fellow researchers will "come to the dataset", instead of that the dataset will be sent to them (MacAvaney et al., in-preperation). In this approach, the sensitive data is kept in a secured environment online, in which researchers could conduct analyses but cannot export the raw information itself. A joint platform by the University of Maryland and NORC at the University of Chicago for example, which provides researchers a secured environment to conduct CL and ML analyses of sensitive data, is the Mental Health Data Enclave (https://enclave.umd.edu/). However, even in these secured platforms, the privacy of the users may be violated, simply because the increase in the number of researchers who are exposed to users' personal information raises the (hopefully unlikely) chance that someone will take a screen shot of the private information. Additional confidentiality measures may therefore be considered. For example, if possible, researchers are encouraged to conduct a thorough anonymization of the dataset, as is conducted in contemporary workshops that address CL research in clinical psychology (e.g., Zirikly et al., 2019), and/or share only aggregated statistics of the data (e.g., list of words used by the at-risk group as a whole, rather than by the individual), which will minimize the risk for privacy violations.

Importantly, the risk for privacy violation does not stop at the raw data level. Sharing the computational representations of the raw data (e.g., the DNN vectors that represent the text or even the trained parameters of DNN models) is also problematic. Although somewhat complicated, adversarial attackers can decode these representations/parameters and recover original texts that served as input examples at the training phase of the model, or other sensitive information about the users for which the model makes predictions. Recent studies on DNN models have illustrated how this data extraction process may reveal private information about the authors of training texts or those for which the model makes predictions, including their names, contact information, and sociodemographic background (Carlini et al., 2020; Coavoux et al., 2018; Li et al., 2018). Additional information regarding the psychosocial background of the authors may be also compromised, since these models, as mentioned above, may capture this sensitive information even when it does not appear in the text in an explicit manner (section 2.1). Indeed, contemporary CL researchers are currently developing defense methods that aim to obscure personal information from texts, thus preventing potential attackers from abusing it, however, further progress in this ethical issue is still needed.

Since both sides of the resource sharing dilemma are important – maintaining the privacy of people on the one hand, and advancing CL research on the other hand – we call for open discussions and institution-based solutions for this dilemma (e.g., by the Association for Psychological Science and other community level institutions such as local associations and even specific universities and research centers). Indeed, general guidelines for AI applications are published from time to time by research institutions and public organizations (Jobin et al., 2019), however, it is essential that a consensual 'best practice' protocol will be developed specifically for resource sharing in suicide research. This future protocol, for example, may support the above described initiative to upload datasets and models to designated servers (MacAvaney et al., in-preperation). A password-protected access to these servers may be decided to be given only to researchers from recognized academic institutions, upon signing a non-disclosure form. Further discussions should be held whether physical infrastructures, such as the proposed secured server should be operated and maintained by researchers, at the professional community level. As new technology constantly emerges, these ongoing ethical discussions are expected to be quite intricate and technical, thus requiring multidisciplinary expertise in a number of fields, including for example, psychology, computer science, computer security, medicine, law, and philosophy. It is also possible that the complexity and sensitivity of these discussions would require us to conduct them (and determine the most reasonable solutions) at the governmental and the legislative level, to make sure that the advancements in CL research will not come at the expense of basic human rights (Dawson et al., 2019).

4.2. The autonomy vs. the safety of individuals at risk for suicide. A second critical concern relates to the safety of suicidal individuals. While the overall goal of CL-suicide applications is virtuous – to promote suicide prevention – the actual process poses a significant ethical obstacle: How to keep the individuals' safety without compromising their privacy (Gibson et al., 2013; Lakeman & Fitzgerald, 2009). While some mental health researchers, including ourselves, are prioritizing lifesaving over most considerations, others might sanctify the right to privacy. Similarly, some people may believe that the use of suicide detection tools should become a legitimate practice in the same way that the use of street cameras have become a common practice in the battle against crime, while others might distinguish between crime and suicide and argue that adults are free to choose whether to live or die.

The complexity of this issue, including its philosophical and social aspects, extends beyond the scope of the current article. Yet, we do wish to propose principal ethical guidelines for research and practice in CL-based suicide prevention. While working on our own research projects (Ophir, Sisso, et al., 2020; Ophir, Tikochinski, et al., 2020), we faced a similar dilemma, which led us to initiate a consortium of experts in suicide and cyber-psychology who discussed the advantages and risks in suicide research online. The discussions revolved around the safety and privacy of research participants who are recruited from crowdsourcing platforms online (which means that their place of residence is not limited to the physical location of the researcher), and yielded a step-by-step protocol for a secured suicide research online (Ophir et al., 2021). Briefly, this protocol proposes that at-risk participants will receive a designated letter in which the researchers encourage them to seek professional help. This letter should be sent to the participants as soon as they complete their research tasks and should include contact information of *local* mental health services. Notably, since contacting participants after the research may be experienced by some of them as intrusion or violation of their privacy, researchers should explicitly inform their participants, at the very beginning of the study (e.g., in the consent form), that they might be contacted by the researcher, if their data would indicate that they are at risk.

Of course, this last recommendation has pros and cons. On the one hand, sending a designated letter to participants may be more suitable to suicide research online than to face-to-face research, since online crowdsourcing platforms usually assign serial numbers to their users and conceal any identifying information. This means that researchers can send the designated letter through the website interface without the need to ask participants for their contact details and without compromising their privacy, thus avoiding the usual trade-off between privacy and safety (Ophir et al., 2021). On the other hand, there is a unique challenge in suicide research online to provide contact information of local mental health services. Indeed, the emergence of the Internet and the social media allowed us to reach large, diverged, and even hard-to-reach populations (Ophir, Sisso, et al., 2020), however these populations might be spread over wide and distant geographical locations, for which researchers do not have specific information regarding local mental health services. Some countries may have wide-ranging 24/7 emergency services, such as the National Suicide Prevention Lifeline or the Crisis Text Line in the US, but it is the researchers' responsibility to collect additional contact details of mental health services that are located at the same geographical region of their participants (Ophir et al., 2021). This recommendation is especially relevant to at-risk individuals who shy away from anonymous and large services and feel more comfortable in face-to-face interactions in small clinics.

Another proposal for suicide researchers who conduct studies online is to leverage the Internet-based communication to provide their at-risk participants with first mental aid online, in a similar way to the support that is provided by designated suicide 'hot lines'. Indeed, a mental health support by trained clinicians may be expensive. However, researchers may consider applying automated intervention platforms that have been demonstrated to increase further help seeking among individuals at risk (Jaroszewski et al., 2019). Either way, it is our view that, as clinicians and/or suicide researchers, we have a unique responsibility to convey a message of hope to our patients/research participants, insist that other and better solutions exist to any life crisis, and encourage them to seek further professional help (Klomek, 2020).

Real-life settings create even more complicated ethical dilemmas: Does the value of lifesaving (suicide prevention) allow us to use automated tools to monitor individuals at risk, even without their explicit permissions? Should we only look at aggregated socio-geographical trends or can we search for individual risks? And how and when should researchers and clinicians respond to individuals at risk who will be detected through CL-based tools? Possible responses may range from providing general information regarding suicide and mental health services to a more specific suggestion to develop a personal safety plan and/or receive psychosocial help from a designated automatic or human-based service. In extreme cases, operators of such applications may even consider involving governmental officials, such as local police departments or social services. What then should be the best future clinical and ethical practices and who has the power to decide which response should be carried out?

Of course, the answers to these questions vary between persons and between societies/cultures. However, it is our stand that we (the scientific community) should advocate that the same ethical norms that are customary in research settings will be applied also in real life settings, that is: Developers and clinicians should prioritize lifesaving while trying their best to minimize the violation of users' privacy and/or autonomy (Jobin et al., 2019). Whether these users are parents to adolescents at-risk, consenting adults, or governmental officials, operators of such sensitive tools should be transparent regarding the information they collect about costumers, obtain clearly written informed consent forms, and allow clients to opt out from the service, whenever they choose to (Paul & Dredze, 2017; Resnik et al., 2020).

4.3. Benefits of real-life monitoring applications vs. risks for intentional data abuse.

Indeed, as thoroughly discussed in the current article, future real-life monitoring applications may have a substantial positive impact on the mental health of many. However, it should be acknowledged that once the technology will be deployed among large populations, the risk for intentional data abuse would increase. This is because the sensitive information about people that is gathered by suicide detection tools is of great value to some parties of interest. It is possible, for example, that pharmaceutical companies would try to utilize this information for micro-targeting purposes – deliberately directing their marketing efforts at users who were identified by automated tools as individuals at risk. Similarly, employers may try to take advantage of these tools and "spy" on potential candidates, labeling them as 'mentally unstable' and rejecting their job applications without further (and fare) screening practices. Researchers, clinicians, and software companies should therefore acknowledge the potential hazards that accompany these powerful tools and consistently evaluate their goals and procedures, while maintaining an open and transparent discourse with the scientific community, the mental health system, and the general public (Paul & Dredze, 2017). The scientific community and public officials should insist that developers and operators of such applications would avoid problematic conflicts of interest and do whatever in their power to prevent the abuse of personal information for financial or political profit.

4.4. The wish to utilize advanced CL systems vs. our limited knowledge regarding their internal mechanism. Finally, the last dilemma concerns our understandable wish to use the powerful tools of CL for positive purposes (e.g., suicide detection) although we have quite limited knowledge regarding their internal mechanism. While the methodological aspects of this "black box" problem are discussed in detail in the next section (section 5.1), here we present a specific example for this problem that is highly relevant to CL-based research and has significant ethical consequences. Recent studies revealed that CL models capture implicit social biases. Since most language representation models (e.g., BERT) are pre-trained on large corpora from general domains (e.g. Wikipedia), they ought to absorb unfounded prejudice notions and stereotypes about different social groups, minorities and majorities alike. Unfortunately, in many cases, these stereotypes include mental health related beliefs that a certain group of people is predisposed to a certain mental illness. Thus, not only do these implicit language biases endanger the accuracy of CL-based predictions, but they can actually also cause harm to individuals, especially if these individuals already belong to a discriminated ethnical or sexual minority (for a comprehensive discussion of this issue, which is known in CL research as 'Bias', see in: Shah et al., 2019). Researchers should therefore be aware of such potential biases and consider utilizing the interpretability frameworks for CL models described above (see section 2.3), which could uncover implicit stereotypes that influence the model's predictions.

5. Challenges and practical recommendations

Regardless of ethical issues, readers that are less familiar with software and algorithms should keep in mind that "machine learning is not magic" (Domingos, 2012) and that its relevance for everyday clinical practice is still debatable. In fact, prominent critiques warn from overreliance on ML methods in suicide research (Siddaway et al., 2020). A recent critical commentary, for example, published in this journal (Jacobucci et al., 2021), challenged the prediction superiority of ML methods over more traditional methods, such as standard logistic regression, and mention a few examples in which both methods achieved similar results (e.g., van Mens et al., 2020). Moreover, the authors have evidenced artificially inflated prediction performances in some machine learning methods (specifically when researchers paired optimism-corrected bootstrap with random forests, instead of using internal validation methods, such as k-fold cross-validation). Indeed, these critiques are not suggesting that machine learning strategies should be neglected altogether, however they do raise principal concerns regarding the generalizability and interpretability of ML-based prediction models. The complexity of ML models according to these critics (and we tend to agree), can be a double-edged sword: On the one hand, this complexity allows high prediction performances in a given task on a given dataset. On the other hand it limits the replication of these models in other settings (and therefore their generalization), and challenges our ability to understand their mechanism (Siddaway et al., 2020). Indeed, several aspects of ML research specifically address these challenges, including: (1) at least two methodological phases (i.e., training and test), (2) relying on strict experimental procedures, such as K-fold cross-validation and statistically sound evaluations (Dror et al., 2018), (3) evaluating the resulted models on multiple datasets and in various settings, and (4) applying interpretational frameworks (section 2.3). However, in this section we wish to dive deeper, into the most dominant challenges in CL-based suicide research and provide practical methodological recommendations, so researchers could dismantle the many mines that are scattered throughout the field.

5.1. Increasing interpretability. Like most other data-driven methodologies, CL-based prediction models are usually not constructed from datasets that were collected in randomized control trials. In fact, they are rarely based on top-down, theory-driven hypotheses testing (De Choudhury & Kiciman 2018; Ernala et al., 2019). Instead, CL models rely heavily on correlational, bottom-up analyses of observational datasets, which challenge the interpretability of their predictions (Feder et al., in press). The complexity of these models and the multiplicity of their input variables, as mentioned above, add another obstacle, which limits their interpretability. In a way, CL models can be viewed as "black boxes" (Alishahi et al., 2019) that are impressive in their ability to provide accurate predictions, but are also leaving the operator of the "box" with very limited ability to explain how these models work. In other words, it is hard to pinpoint what were the specific indications that allowed the machine to make its predictions. Moreover, even when researchers aim to identify these specific indications (e.g., through the usage of interpretational techniques), they might derive inaccurate conclusions because predictors of suicide risk are typically characterized with multicollinearity (e.g., depression and anxiety) that challenges their separation from one another. Considering these characteristics, our ability to derive clear conclusions from CL-based findings is limited (Chancellor & De Choudhury, 2020).

To address this challenge, we propose three principal solutions: First, researchers are advised to enrich data-driven 'bottom-up' studies with specific theory-driven insights and risk factors from the large body of research on suicide and its related mental disorders (Zaman, 2021). By doing so, researchers would leverage the distinct benefit of the traditional, top-down approach in suicide prevention, and be able to formulate specific a-priori hypotheses, which could, if confirmed, increase the explainability of CL prediction models. In our previous work for example, we hypothesized that a multi-task CL model, which was designed to predict a hierarchy of theory-driven risk factors (i.e., psychiatric disorders, psychosocial risk factors, and personality traits) alongside the risk for suicide, could yield improved predictions compared with single-task models, which aim to predict suicide risk only (Ophir, Tikochinski, et al., 2020). The confirmation of this hypothesis contributes (to a certain extent) to the explainability of the multi-task model because it suggests that it extracts a multifaceted psychosocial profile for each user – a profile that is theoretically linked to suicide ideation and behaviors – and produces its (improved) predictions regarding the suicidal risk of the user, based on this rich profile.

Second, researchers should consider utilizing interpretability frameworks for CL models (see section 2.3) and especially methodologies that tease apart causation from mere correlation. Together with the previous recommendation to leverage the accumulating (top-down) knowledge, the application of such frameworks may facilitate the appraisal of explanations regarding the generalizability and the unique predictive role of specific features in the model. Notably, these frameworks are particularly suitable for: (1) dealing with the statistical problem described above of multicollinearity, (2) combating the implicit social stereotypes described in the section on ethics (section 4.3), and for (3) exploring top-down hypotheses regarding the effects of relatively abstract concepts, such as internal emotional states or psychological disturbances. Finally, we recommend the usage of qualitative human interpretations as a complimentary analysis to the CL ones. For example, researchers may carefully read small portions of the texts that were inserted as an input to the computational model, and manually search for meaningful language signals in texts that were generated by suicidal individuals. These signals may then be compared with the patterns identified by the interpretational techniques (2.3) and lend convergent support to, or a refutation of, a given a-priori hypothesis regarding the theoretical construct of suicide. Eventually, these human and machine interpretational techniques can "brighten the black box" of CL prediction models and deepen our understanding of the mental state of people suffering from suicide ideation.

5.2. Increasing generalizability. Another well-recognized problem of CL research (and ML research in general) is the difficulty to generalize findings form one dataset to another. By its nature, every dataset reflects a specific and somewhat unique image of the world. Each dataset consists of specific types of texts (e.g. medical records vs. social media), specific labeling scheme (e.g. determining the risk through online 'warning signs' vs. screening tools), a specific number of training examples, and specific types of inputs (e.g., texts vs. images). This means that a model that was trained on one dataset (e.g., a relatively small sample of highly reliable medical records) may not yield accurate results when applied to another dataset (e.g., a large sample of labeled posts from Facebook). Moreover, even when both datasets represent a similar source (e.g., social media), the generalizability from one set (e.g., a set from Facebook) to another (e.g., a set from Instagram) may not be trivial. Consider for example, populations of adolescents or young adults, who are considered a relatively vulnerable age group at risk for suicide (Varnik, 2012). Over the years, young social media users have shifted from the more text-oriented platform of Facebook to other social networks that emphasize more visual and/or short contents, such as Instagram, You Tube, and Snapchat (Pew Research Center, 2018). Their language usage online may also differ significantly from older users' language, since many of them use slang words (friending, trolling), repetitions (Yesssss!!!), abbreviations (LOL), onomatopoetic laughter (Hahahaha), discourse markers (#), kisses/hugs signs (XOXO) and emoji (😎) (Hilte et al., 2018). Thus, CL-based models that were trained on adult language in Facebook may yield poor results for adolescent language in Instagram.

In order to establish a broad coverage of real-world scenarios, much efforts are directed today at domain adaptation research, that is, developing methods that can help adapt a model that was trained with data from one textual domain so that it can produce accurate predictions on data from other domains (e.g., Ziser & Reichart, 2017; Ziser & Reichart, 2018). This scientific effort requires of course that scientists will share their research material with the scientific community – a practice that has complicated ethical aspects (section 4.1). Researchers are therefore encouraged to keep developing secured methods to share their data (e.g., through secured platforms), while maintaining an open discourse regarding this ethical dilemma, both within the scientific community and outside the community, at the public and governmental level (for further discussion of this issue, see section 4.1). This is because only by allowing access to fellow researchers we could engage in domain adaptation research and increase the generalizability of prediction models across various datasets and age groups.

We also join previous recommendations that researchers will ensure the generalizability of their model through internal validation methods, such as k-fold cross-validation (Jacobucci et al., 2021). Then, upon completion of a given study, researchers are advised to be transparent and provide the scientific community with detailed information regarding their models and their related parameters (whether within the scientific publication or as an external open source material) (Dodge et al., 2019; Siddaway et al., 2020). This practice could allow other researchers to conduct replication studies, even in cases where the actual data sharing is restricted due to privacy considerations.

5.3. Establishing high quality ground truth labels. Another major challenge concerns the reliability and construct validity of the data that is being used for the establishment of a 'ground truth' criterion for suicide risk (or for any other mental health condition, for that matter) (Chancellor & De Choudhury, 2020). On the one hand, using ground truth labels that rely on human annotations of (online) textual data may be inappropriate for measuring suicide risk and may also miss the large populations that refrain from articulating their emotional pain in an explicit manner (Ophir, 2017). On the other hand, ground truth labels that rely on external self-reported psychological information (e.g., self-report questionnaires) may suffer from poor reliability, especially if this information originates from the popular crowdsourcing platforms (e.g., Amazon's Mechanical Turk) that allow researchers to reach large and diverge samples of research participants with a relatively little investment of time and research budget.

To overcome this obstacle, researchers are advised to make efforts to collect external psychological ground truths (see section 1.2), but if they choose to collect this information from crowdsourcing platforms, they should apply rigid quality measures that could reduce the rates of false/fake responses (Ophir, Tikochinski, et al., 2020). This is because there is growing evidence that crowdsourcing research samples include large proportions of inattentive, and even bogus participants ('Bots') (Chandler et al., 2020; Ophir, Sisso, et al., 2020). Moreover, there is a significant concern that even authenticated users of such platforms are not representative of the general population. Compared with representative national surveys, crowdsourcing platforms seem to include substantially higher levels of psychopathologies, and especially of depression (Arditte et al., 2016; Bunge et al., 2018; Chandler & Shapiro, 2016; McCredie & Morey, 2018; Ophir, Sisso, et al., 2020), thus limiting the generalizability of prediction models that are based on data from such platforms. Researchers are therefore advised to apply strict data quality protocols and implement multiple validity measures at the early, data-collection stage of their research (Chancellor & De Choudhury, 2020; Chandler et al., 2020; Ophir, Sisso, et al., 2020). We also recommend that researchers will try to recruit participants from multiple sources and test the prediction performances of their CL models in different settings.

Aside from the general validity issue that characterizes most self-report measures, there is a unique challenge in suicide assessment that concerns the wide heterogeneity of suicide ideation/behaviors (see also in section 1.3) and the difficulty to detect highly lethal/dangerous suicide attempts. On the one hand, studies that aim to detect pure suicide ideation may have relatively low predictive value for detecting imminent suicide risk (McHugh & Large, 2020). This is because, like all other risk factors (see the introduction), suicide ideation alone is not sufficient to predict serious suicide attempts. Many people may experience suicidal thoughts but only few of them act on these thoughts (of which only a small percentage die, following these attempts). On the other hand, studies that aim to predict only actual suicide deaths as a gold standard criterion, may leave the CL algorithm with too little positive examples to learn from. This is because actual deaths are extremely rare in random samples, despite the fact that suicide is one of the most prevalent causes of death.

Although the detection of suicide ideation has a merit on its own, as it could facilitate the distribution of psychosocial support to individuals in need, our recommendation to researchers is to integrate data regarding all levels of suicide risk, including actual deaths and near fatal suicide attempts (i.e., a subgroup of individuals who attempted suicide but did not die). This last subgroup may be somewhat reluctant to share their experience but it could serve as a suitable "proxy" for actual cases of suicide (Douglas et al., 2004). Then, in the construction of the models, and in the analysis and interpretation of the results, researchers are encouraged to make a clear distinction between low risk suicide ideation, high risk suicide plans or behaviors, and near fatal suicide attempts.

It is noted, however, that even if such clear distinctions are made and recruitment efforts are directed at high risk populations, the number of people who would meet the imminent risk criterion is expected to be relatively small. This creates a "class imbalance" problem, a well-recognized problem in ML research that refers to cases where only a small fraction of the dataset belongs to important classes that are of interest for the designer of the algorithm. Researchers are therefore advised to implement appropriate ML strategies that take into account this class imbalance problem (e.g., Haixiang et al., 2017), which is an inherent problem to the specific classification task of suicide detection.

5.4. Fine-tuning and calibrating CL classifiers. Although most contemporary language representation models are well trained, they may still need to be adjusted to specific domains of interest (e.g., Facebook posts), to allow the construction of an effective supervised classifier. Indeed, CL-suicide researchers do not need to collect billions of unlabeled examples (i.e. free text), as was done to train the static and the contextualized embedding models (section 2.1). However, they might still need to collect hundreds-to-thousands of *labeled* textual examples, in order to construct supervised classifiers that could distinguish between different types of psychological profiles. This set of training examples can be quite costly and their collection process can be time consuming, especially when the labels (predictive criterion) constitute expert/laymen annotations of each text (e.g., Zirikly et al., 2019) or high quality psychological/psychiatric information about the authors of the texts (e.g., Coppersmith et al., 2018; Eichstaedt et al., 2018; Ophir, Tikochinski, et al., 2020).

Then, after the construction of the classifier based on this dataset, it is important to ensure that the classifier produces reliable predictions. Most ML models augment their predictions with a confidence score. For example, a classifier that predicts if a person is at risk or not would typically compute the probability of the positive class (that the person *is* at risk) and the probability of the negative class (that the person *is not* at risk) and would output the higher probability class. In such a classifier, the probability of the chosen class also serves as the confidence score of the model. Importantly, a high correlation between this confidence score and the probability that the classifier is indeed correct, indicates that the classifier is well-calibrated. This means that the classifier's confidence scores could be utilized for assessing the actual risk. Positive classifications (that the person *is* at risk) with a confidence score close to 1 would indicate that the risk is highly reliable (that the person is indeed experiencing suicidal ideation) and a probability score close to 0.5, would indicate that the risk is less reliable.

Unfortunately, uncalibrated models are not uncommon in ML studies (Desai & Durrett, 2020; Guo et al., 2017), thus deterring potential operators (e.g., clinicians) from embracing and applying these tools in their daily practice. From the clinician perspective, this can be a practical and ethical dilemma. Whether false positive or false negative, any miscalculation of the real risk for suicide could trigger problematic real-life consequences. Determining that a non-suicidal person is at risk may be followed by unnecessary pharmacological treatments and perhaps even hospitalizations, and determining that a suicidal person is not at risk may stall crucial treatments from individuals in need (Siddaway et al., 2020). Notably however, there are well-researched solutions for this obstacle. When properly constructed, the reliability of CL models can be improved using standard calibration techniques, such as temperature scaling and label smoothing (Desai & Durrett, 2020). Researchers and developers are therefore encouraged to ensure high quality calibration of their CL-based classifiers, which will provide them decent level of confidence in their predictions.

5.5. Considering real-world challenges within inter-disciplinary collaborations. Finally, multiple practical challenges are expected to arise once CL-based research will mature into real-life monitoring applications and other psycho-diagnostic tools. Without a clear understanding of the inner mechanism of computational prediction models (section 5.1), practitioners and policy makers may hesitate to adopt the new technology. Unlike the face-validity of self-report screening questionnaires or the reasonable rationales of indirect psychological assessment tools, the reliance on a "black box" that is said to base its predictions on subtle language signals may be perceived as a far-reaching step by many policy makers and clinicians. Moreover, even if clinicians would agree to integrate CL-based tools in their daily practice, the exact procedures in which these tools will be applied remain somewhat vague and highly challenging (see also section 4.2). Consider, for example, the application of automated suicide screening tools. How would operators of such tools receive access to textual data of potential users and who will determine the many optional thresholds for the various types of suicide risk? After all, most CL models do not provide operators with binary (yes/no) answers, but with probabilities (i.e., not certainties) that are supposed to reflect a suicide risk.

Although we believe that these questions should not stop us from keep investigating the potential of CL in suicide prevention, we concur with previous cautious statements regarding the hazards of this field (Siddaway et al., 2020) and call for a careful implementation of CL tools alongside, and not instead of, the existing traditional methods for suicide screening, psychological assessments, and therapeutic interventions. It is also recommended to form genuine, multi-disciplinary collaborations between computational and psychological scientists, as well with philosophers, legal scholars, and clinicians who deal with suicidal patients daily. Close collaboration between researchers and clinicians may also reduce the expected concerns and barriers of the latter and encourage clinicians to acquire a basic acquaintance with CL frameworks.

Summary

Our last recommendation from the previous section, which emphasizes the importance of interdisciplinary collaborations, seems to be a key component in CL-based suicide prevention (Ophir, Tikochinski, et al., 2020; Resnik et al., 2020). Indeed, further research is crucially needed to keep developing this field while addressing its inherent ethical and methodological challenges (sections 4-5). However, we believe that interdisciplinary research teams that will integrate the state-of-the-art DNN technologies in their CL-suicide research could significantly enhance our ability to detect and understand suicide ideation and behaviors (sections 1-3).

Notably, as illustrated in this article, a growing body of literature suggests that CL tools can be utilized for constructing high-quality suicide predictions, reaching out to large populations by analyzing their informal, everyday language on the various social media platforms, and tracking their suicide ideation and behaviors in real time (section 1). Moreover, the application of CL strategies, such as word embeddings, topic modeling, or interpretation frameworks, might deepen our theoretical understanding of suicide behaviors and their related mental health disorders such as depression (section 2). Finally, it is expected that CL would facilitate the personalized medicine approach through the application of data-driven methodologies that would target principal issues in psychological assessment, such as the subjectivity of self-report information, the heterogeneity within the mental health diagnoses, and the difficulty to conduct indirect, yet psychometrically validated psychological assessments (section 3).

Of course, all these promising futuristic developments should be carefully and openly discussed by scientists and clinicians, ethics experts and policy makers, since any progress in this field is bound to trigger serious ethical dilemmas. Aside from the highly discussed issue of privacy in the digital era, CL-suicide prevention involves complicated dilemmas, such as how to keep peoples' safety with a minimum violation of their privacy and how to ensure that adversarial attackers and financial/political parties will not abuse people's sensitive data for non-scientific or non-clinical purposes (section 4). Additional limitations are the inherent methodological obstacles in CL research, such as the difficulty to interpret the "black box" of DNN models and to generalize them to various populations and settings (section 5).

In the current article, we nevertheless strived to deliver an optimistic message and propose a handful of practical solutions and principle guidelines, which could help researchers and clinicians overcome these complicated limitations. In the ethic section for example, we presented the newly developed secured environment, in which researchers "come to the dataset", instead of the other way around, and reviewed several ways to maintain the safety of individuals at risk. In the methodological limitations section for example, we discussed how interpretational frameworks can be used to "brighten the black box" of CL models and what can be done to ensure the validity and generalizability of these models. In this way, the current wide-scope overview article may serve as an informative, yet relatively easy to read, travel guide to CL in suicide prevention for researchers, clinicians, and policy makers, who may be less familiar with current advancements in AI technologies. It is our belief that future (interdisciplinary) research teams and practitioners could utilize the CL methods presented in this article to draw fascinating new research paths, which will have a substantial impact on the way we perceive, detect, and treat suicide ideation and behaviors.

References

Aharoni, R., & Goldberg, Y. (2020). Unsupervised Domain Clusters in Pretrained Language Models. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7747-7763.

Alishahi, A., Chrupała, G., & Linzen, T. (2019). Analyzing and interpreting neural networks for NLP: A report on the first BlackboxNLP workshop. *Natural Language Engineering*, *25*(4), 543-557.

Alpaydin, E. (2020). *Introduction to machine learning*. MIT press.

American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders (DSM-5®)*. American Psychiatric Pub.

Arditte, K. A., Çek, D., Shaw, A. M., & Timpano, K. R. (2016). The importance of assessing clinical phenomena in Mechanical Turk research. *Psychological Assessment*, *28*(6), 684.

Aubin, H.-J., Berlin, I., & Kornreich, C. (2013). The evolutionary puzzle of suicide. *International journal of environmental research and public health*, *10*(12), 6873-6886.

Banerjee, D., Kosagisharaf, J. R., & Rao, T. S. S. (2020). 'The dual pandemic'of suicide and COVID-19: A biopsychosocial narrative of risks and prevention. *Psychiatry research*, 113577.

Beck, A. T. (1991). Cognitive therapy: A 30-year retrospective. *American psychologist*, *46*(4), 368.

Belinkov, Y. (2021). Probing Classifiers: Promises, Shortcomings, and Alternatives. *arXiv preprint arXiv:2102.12452*.

Bernert, R. A., Hilberg, A. M., Melia, R., Kim, J. P., Shah, N. H., & Abnousi, F. (2020). Artificial intelligence and suicide prevention: a systematic review of machine learning investigations. *International journal of environmental research and public health*, *17*(16), 5929.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, *3*, 993-1022.

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., & Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 4356-4364.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., & Askell, A. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Bruffaerts, R., Demyttenaere, K., Hwang, I., Chiu, W. T., Sampson, N., Kessler, R. C., Alonso, J., Borges, G., de Girolamo, G., & de Graaf, R. (2011). Treatment of suicidal people around the world. *The British Journal of Psychiatry*, *199*(1), 64-70.

Bunge, E., Cook, H. M., Bond, M., Williamson, R. E., Cano, M., Barrera, A. Z., Leykin, Y., & Muñoz, R. F. (2018). Comparing Amazon Mechanical Turk with unpaid internet resources in online clinical trials. *Internet Interventions*, *12*, 68-73. https://doi.org/https://doi.org/10.1016/j.invent.2018.04.001

Burke, T. A., Ammerman, B. A., & Jacobucci, R. (2019). The use of machine learning in the study of suicidal and non-suicidal self-injurious thoughts and behaviors: A systematic review. *Journal of affective disorders*, *245*, 869-884.

Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., & Erlingsson, U. (2020). Extracting Training Data from Large Language Models. *arXiv preprint arXiv:2012.07805*.

Chancellor, S., & De Choudhury, M. (2020). Methods in predictive techniques for mental health status on social media: a critical review. *npj Digital Medicine*, *3*(1), 43. https://doi.org/10.1038/s41746-020-0233-7

Chandler, J., & Shapiro, D. (2016). Conducting clinical research using crowdsourced convenience samples. *Annual Review of Clinical Psychology*, *12*.

Chandler, J., Sisso, I., & Shapiro, D. (2020). Participant carelessness and fraud: Consequences for clinical research and potential solutions. *Journal of Abnormal Psychology*, *129*(1), 49.

Clark, A., Fox, C., & Lappin, S. (2013). *The handbook of computational linguistics and natural language processing*. John Wiley & Sons.

Clark, D. A., & Beck, A. T. (2010). Cognitive theory and therapy of anxiety and depression: convergence with neurobiological findings. *Trends in cognitive sciences*, *14*(9), 418-424.

Coavoux, M., Narayan, S., & Cohen, S. B. (2018). Privacy-preserving Neural Representations of Text. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 1-10.

Collins, F. S., & Varmus, H. (2015). A new initiative on precision medicine. *New England journal of medicine*, *372*(9), 793-795.

Conneau, A., Schwenk, H., Cun, Y. L., & Barrault, L. (2017). Very deep convolutional networks for text classification. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 1107-1116.

Coppersmith, G., Leary, R., Crutchley, P., & Fine, A. (2018). Natural Language Processing of Social Media as Screening for Suicide Risk. *Biomedical Informatics Insights*, *10*, 1178222618792860. https://doi.org/10.1177/1178222618792860

Dawson, D., Schleiger, E., Horton, J., McLaughlin, J., Robinson, C., Quezada, G., Scowcroft, J., & Hajkowicz, S. (2019). Artificial intelligence: Australia’s ethics framework.

De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. (2013). Predicting Depression via Social Media. *ICWSM*, *13*, 1-10.

De Choudhury, M., & Kiciman , E. (2018). Integrating Online and Offline Data in Complex, Sensitive Problem Domains: Experiences from Mental Health. *Menlo Park - Association for the Advancement of Artificial Intelligence*.

De Choudhury, M., Kiciman, E., Dredze, M., Coppersmith, G., & Kumar, M. (2016). Discovering shifts to suicidal ideation from mental health content in social media. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2098-2110.

Desai, S., & Durrett, G. (2020). Calibration of Pre-trained Transformers. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 295-302.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dodge, J., Gururangan, S., Card, D., Schwartz, R., & Smith, N. A. (2019). Show Your Work: Improved Reporting of Experimental Results. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2185-2194.

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM, 55*(10), 78-87.

Douglas, J., Cooper, J., Amos, T., Webb, R., Guthrie, E., & Appleby, L. (2004). "Near-fatal" deliberate self-harm: characteristics, prevention and implications for the prevention of suicide. *Journal of affective disorders*, *79*(1-3), 263-268.

Dror, R., Baumer, G., Shlomov, S., & Reichart, R. (2018). The hitchhiker's guide to testing statistical significance in natural language processing. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1383-1392.

Eichstaedt, J. C., Smith, R. J., Merchant, R. M., Ungar, L. H., Crutchley, P., Preoţiuc-Pietro, D., Asch, D. A., & Schwartz, H. A. (2018). Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences*, *115*(44), 11203-11208.

Engel, G. L. (1977). The need for a new medical model: a challenge for biomedicine. *Science*, *196*(4286), 129-136.

Engel, G. L. (1981). The clinical application of the biopsychosocial model. *The Journal of Medicine and Philosophy: A Forum for Bioethics and Philosophy of Medicine*, *6*, 101-124.

Ernala, S. K., Birnbaum, M. L., Candan, K. A., Rizvi, A. F., Sterling, W. A., Kane, J. M., & De Choudhury, M. (2019). Methodological Gaps in Predicting Mental Health States from Social Media: Triangulating Diagnostic Signals. *Proceedings of the 2019 CHI conference on human factors in computing systems*, 1-16.

Evans-Lacko, S., & Knapp, M. (2016). Global patterns of workplace productivity for people with depression: absenteeism and presenteeism costs across eight diverse countries. *Social Psychiatry and Psychiatric Epidemiology*, *51*(11), 1525-1537.

Fasold, R. W. (1990). *The sociolinguistics of language* (Vol. 2). Blackwell Pub.

Feder, A., Oved, N., Shalit, U., & Reichart, R. (in press). CausaLM: Causal Model Explanation Through Counterfactual Language Models. *Computational Linguistics*, *Available at https://arxiv.org/abs/2005.13407*.

Franklin, J. C., Ribeiro, J. D., Fox, K. R., Bentley, K. H., Kleiman, E. M., Huang, X., Musacchio, K. M., Jaroszewski, A. C., Chang, B. P., & Nock, M. K. (2017). Risk factors for suicidal thoughts and behaviors: a meta-analysis of 50 years of research. *Psychological bulletin*, *143*(2), 187.

Fried, E. I. (2017). The 52 symptoms of major depression: Lack of content overlap among seven common depression scales. *Journal of Affective Disorders*, *208*, 191-197. https://doi.org/https://doi.org/10.1016/j.jad.2016.10.019

Fried, E. I., & Nesse, R. M. (2015). Depression is not a consistent syndrome: an investigation of unique symptom patterns in the STAR* D study. *Journal of affective disorders*, *172*, 96-102.

Galynker, I., Yaseen, Z. S., Cohen, A., Benhamou, O., Hawes, M., & Briggs, J. (2017). Prediction of suicidal behavior in high risk psychiatric patients using an assessment of acute suicidal state: The suicide crisis inventory. *Depression and Anxiety*, *34*(2), 147-158. https://doi.org/10.1002/da.22559

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, *115*(16), E3635-E3644.

Gibson, S., Benson, O., & Brand, S. L. (2013). Talking about suicide: Confidentiality and anonymity in qualitative research. *Nursing Ethics*, *20*(1), 18-29.

Giuntini, F. T., Cazzolato, M. T., dos Reis, M. d. J. D., Campbell, A. T., Traina, A. J. M., & Ueyama, J. (2020). A review on recognizing depression in social networks: challenges and opportunities. *Journal of Ambient Intelligence and Humanized Computing*, 1-17.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

Groth-Marnat, G. (2009). *Handbook of psychological assessment*. John Wiley & Sons.

Gunnell, D., Appleby, L., Arensman, E., Hawton, K., John, A., Kapur, N., Khan, M., O'Connor, R. C., Pirkis, J., & Caine, E. D. (2020). Suicide risk and prevention during the COVID-19 pandemic. *The Lancet Psychiatry*, *7*(6), 468-471.

Guntuku, S. C., Yaden, D. B., Kern, M. L., Ungar, L. H., & Eichstaedt, J. C. (2017). Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences*, *18*, 43-49.

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. *International Conference on Machine Learning*, 1321-1330.

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, *73*, 220-239.

Hawton, K., & van Heeringen, K. (2009). Suicide. *The Lancet*, *373*(9672), 1372-1381. https://doi.org/https://doi.org/10.1016/S0140-6736(09)60372-X

Hedegaard, H., Curtin, S. C., & Warner, M. (2018). *Suicide rates in the United States continue to increase*. Hyattsville, MD: US Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics. .

Hilte, L., Vandekerckhove, R., & Daelemans, W. (2018). Social media writing and social class: A correlational analysis of adolescent CMC and social background. *International Journal of Society, Culture & Language*, *6*(2), 73-89.

Hitchcock, P., Fried, E. I., & Frank, M. (2021). Computational Psychiatry Needs Time and Context. In: PsyArXiv.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735-1780.

Horowitz, L. M., & Ballard, E. D. (2009). Suicide screening in schools, primary care and emergency departments. *Current opinion in pediatrics*, *21*(5), 620.

Hyman, S. E. (2010). The diagnosis of mental disorders: the problem of reification. *Annual review of clinical psychology*, *6*, 155-179.

Isaak, J., & Hanna, M. J. (2018). User Data Privacy: Facebook, Cambridge Analytica, and Privacy Protection. *Computer*, *51*(8), 56-59. https://doi.org/10.1109/MC.2018.3191268

Jacobucci, R., Littlefield, A. K., Millner, A. J., Kleiman, E. M., & Steinley, D. (2021). Evidence of Inflated Prediction Performance: A Commentary on Machine Learning and Suicide Research. *Clinical Psychological Science*, 2167702620954216.

Jain, S., & Wallace, B. C. (2019). Attention is not Explanation. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 3543-3556.

Jameson, J. L., & Longo, D. L. (2015). Precision medicine—personalized, problematic, and promising. *Obstetrical & gynecological survey*, *70*(10), 612-614.

Jaroszewski, A. C., Morris, R. R., & Nock, M. K. (2019). Randomized controlled trial of an online machine learning-driven risk assessment and intervention platform for increasing the use of crisis services. *Journal of consulting and clinical psychology*, *87*(4), 370.

Ji, S., Pan, S., Li, X., Cambria, E., Long, G., & Huang, Z. (2020). Suicidal ideation detection: A review of machine learning methods and applications. *IEEE Transactions on Computational Social Systems*.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389-399. https://doi.org/10.1038/s42256-019-0088-2

Kleiman, E. M., Turner, B. J., Fedor, S., Beale, E. E., Huffman, J. C., & Nock, M. K. (2017). Examination of real-time fluctuations in suicidal ideation and its risk factors: Results from two ecological momentary assessment studies. *Journal of abnormal psychology*, *126*(6), 726.

Klomek, A. B. (2020). Suicide prevention during the COVID-19 outbreak. *The Lancet Psychiatry*, *7*(5), 390.

Kosinski, M., Matz, S. C., Gosling, S. D., Popov, V., & Stillwell, D. (2015). Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines. *American Psychologist*, *70*(6), 543.

Kramer, A. D. I., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, *111*(24), 8788. https://doi.org/10.1073/pnas.1320040111

Lakeman, R., & Fitzgerald, M. (2009). Ethical suicide research: a survey of researchers. *International journal of mental health nursing*, *18*(1), 10-17.

Lehman, B. J., David, D. M., & Gruber, J. A. (2017). Rethinking the biopsychosocial model of health: Understanding health as a dynamic system [https://doi.org/10.1111/spc3.12328]. *Social and Personality Psychology Compass*, *11*(8), e12328. https://doi.org/https://doi.org/10.1111/spc3.12328

Levi-Belz, Y., Gvion, Y., & Apter, A. (2019). The psychology of suicide: from research understandings to intervention and treatment. *Frontiers in psychiatry*, *10*, 214.

Li, Y., Baldwin, T., & Cohn, T. (2018). Towards Robust and Privacy-preserving Text Representations. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 25-30.

Lilienfeld, S. O., Wood, J. M., & Garb, H. N. (2001). What's wrong with this picture? *Scientific American*, *284*(5), 80-87.

MacAvaney, S., Mittu, A., Coppersmith, G., & Resnik, P. (in-preperation). Community-level research on suicidality prediction in a secure environment: Overview of the CLPsych 2021 Shared Task. . In: Workshop on Computational Linguistics And Clinical Psychology (CLPsych) at the North American Conference on Computational Linguistics (NAACL).

McCredie, M. N., & Morey, L. C. (2018). Who Are the Turkers? A Characterization of MTurk Workers Using the Personality Assessment Inventory. *Assessment*, 1073191118760709.

McCulloch, G. (2020). *Because internet: Understanding the new rules of language*. Riverhead Books.

McHugh, C. M., & Large, M. M. (2020). Can machine-learning methods really help predict suicide? *Current Opinion in Psychiatry*, *33*(4).

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *arXiv preprint arXiv:1310.4546*.

Mneimneh, Z., Pasek, J., Singh, L., Best, R., Bode, L., Bruch, E., Budak, C., Davis-Kean, P., Donato, K., & Ellison, N. (2021). Data Acquisition, Sampling, and Data Preparation Considerations for Quantitative Social Science Research Using Social Media Data. *PsyArXiv*. *March*, *16*.

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.

Olteanu, A., Castillo, C., Diaz, F., & Kiciman, E. (2019). Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, *2*, 13.

Ophir, Y. (2017). SOS on SNS: Adolescent distress on social network sites. *Computers in Human Behavior*, *68*, 51-55. https://doi.org/http://dx.doi.org/10.1016/j.chb.2016.11.025

Ophir, Y., Amichai-Hamburger, Y., Brunstein-Klomek , A., Levi-Belz, Y., Hadlaczky , G., Yom-Tov, E., & Zalsman , G. (2021). The Ethics of Suicide Research Online: A Consensual Protocol for Crowdsourcing-based Studies on Suicide. *PsyArXiv*. https://doi.org/https://doi.org/10.31234/osf.io/bmuyh

Ophir, Y., Asterhan, C. S. C., & Schwarz, B. B. (2019). The digital footprints of adolescent depression, social rejection and victimization of bullying on Facebook. *Computers in Human Behavior*, *91*, 62-71. https://doi.org/https://doi.org/10.1016/j.chb.2018.09.025

Ophir, Y., Rosenberg, H., Lipshits-Braziler, Y., & Amichai-Hamburger, Y. (2020). “Digital adolescence”: The effects of smartphones and social networking technologies on adolescents’ well-being. In *Online Peer Engagement in Adolescence* (pp. 122-139). Routledge.

Ophir, Y., Sisso, I., Asterhan, C. S. C., Tikochinski, R., & Reichart, R. (2020). The turker blues: Hidden factors behind increased depression rates among Amazon’s Mechanical Turkers. *Clinical Psychological Science*, *8*(1), 65-83.

Ophir, Y., Tikochinski, R., Asterhan, C. S. C., Sisso, I., & Reichart, R. (2020). Deep neural networks detect suicide risk from textual facebook posts. *Scientific Reports*, *10*(1), 16685. https://doi.org/10.1038/s41598-020-73917-0

Oved, N., Feder, A., & Reichart, R. (2020). Predicting In-Game Actions from Interviews of NBA Players. *Computational Linguistics*, *46*(3), 667-712.

Paul, M. J., & Dredze, M. (2017). Social monitoring for public health. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, *9*(5), 1-183.

Paulhus, D. L., & Vazire, S. (2007). The self-report method. In *Handbook of research methods in personality psychology* (Vol. 1, pp. 224-239). Guilford;.

Pearl, J. (2009). Causal inference in statistics: An overview. *Statistics surveys*, *3*, 96-146.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532-1543.

Perna, G., Grassi, M., Caldirola, D., & Nemeroff, C. B. (2018). The revolution of personalized psychiatry: will technology make it happen sooner? *Psychological Medicine*, *48*(5), 705-713. https://doi.org/10.1017/S0033291717002859

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Pew Research Center. (2018). Teens, Social Media & Technology 2018. In: Pew Research Center Internet & Technology. Last retrieved on March 8, 2021, from https://www.pewresearch.org/internet/2018/05/31/teens-social-media-technology-2018/.

Pew Research Center. (2019). Social Media Fact Sheet. In: Pew Research Center Internet & Technology. Last retrieved on March 8, 2021, from http://www.Pewinternet.org/fact-sheet/social-media/.

Posner, K., Brown, G. K., Stanley, B., Brent, D. A., Yershova, K. V., Oquendo, M. A., Currier, G. W., Melvin, G. A., Greenhill, L., & Shen, S. (2011). The Columbia–Suicide Severity Rating Scale: initial validity and internal consistency findings from three multisite studies with adolescents and adults. *American Journal of Psychiatry*, *168*(12), 1266-1277.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, *21*, 1-67.

Resnik, P., Foreman, A., Kuchuk, M., Musacchio Schafer, K., & Pinkham, B. (2020). Naturally occurring language as a source of evidence in suicide prevention. *Suicide and Life-Threatening Behavior*.

Ribeiro, J. D., Huang, X., Fox, K. R., Walsh, C. G., & Linthicum, K. P. (2019). Predicting imminent suicidal thoughts and nonfatal attempts: The role of complexity. *Clinical Psychological Science*, *7*(5), 941-957.

Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, *8*, 842-866.

Roy, A., Nikolitch, K., McGinn, R., Jinah, S., Klement, W., & Kaminsky, Z. A. (2020). A machine learning approach predicts future risk to suicidal ideation from social media data. *npj Digital Medicine*, *3*(1), 78. https://doi.org/10.1038/s41746-020-0287-6

Rude, S., Gortner, E.-M., & Pennebaker, J. (2004). Language use of depressed and depression-vulnerable college students. *Cognition & Emotion*, *18*(8), 1121-1133.

Ruder, S. (2017). An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Schafer, K., Kennedy, G., Gallyer, A., & Resnik, P. (in press). A Direct Comparison of Theory-Driven and Machine Learning Prediction of Suicide: AMeta-Analysis. *PLOS ONE*.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., & Seligman, M. E. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, *8*(9), e73791.

Sejnowski, T. J. (2018). *The deep learning revolution*. Mit Press.

Shah, D., Schwartz, H. A., & Hovy, D. (2019). Predictive biases in natural language processing models: A conceptual framework and overview. *arXiv preprint arXiv:1912.11078*.

Shatte, A. B. R., Hutchinson, D. M., & Teague, S. J. (2019). Machine learning in mental health: a scoping review of methods and applications. *Psychological Medicine*, *49*(9), 1426-1448. https://doi.org/10.1017/S0033291719000151

Siddaway, A. P., Quinlivan, L., Kapur, N., O'Connor, R. C., & de Beurs, D. (2020). Cautions, concerns, and future directions for using machine learning in relation to mental health problems and clinical and forensic risks: A brief comment on “Model complexity improves the prediction of nonsuicidal self-injury”(Fox et al., 2019).

Smith, J. M., Gacono, C. B., Fontan, P., Taylor, E. E., Cunliffe, T. B., & Andronikof, A. (2018). A scientific critique of Rorschach research: Revisiting Exner’s Issues and Methods in Rorschach Research (1995). *Rorschachiana*, *39*(2), 180.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818-2826.

Tadesse, M. M., Lin, H., Xu, B., & Yang, L. (2019). Detection of depression-related posts in reddit social media forum. *IEEE Access*, *7*, 44883-44893.

Thom, R., Hogan, C., & Hazen, E. (2020). Suicide risk screening in the hospital setting: a review of brief validated tools. *Psychosomatics*, *61*(1), 1-7.

Turecki, G., Brent, D. A., Gunnell, D., O’Connor, R. C., Oquendo, M. A., Pirkis, J., & Stanley, B. H. (2019). Suicide and suicide risk. *Nature Reviews Disease Primers*, *5*(1), 74. https://doi.org/10.1038/s41572-019-0121-0

van Mens, K., de Schepper, C. W. M., Wijnen, B., Koldijk, S. J., Schnack, H., de Looff, P., Lokkerbol, J., Wetherall, K., Cleare, S., C O'Connor, R., & de Beurs, D. (2020). Predicting future suicidal behaviour in young adults, with different machine learning techniques: A population-based longitudinal study. *Journal of Affective Disorders*, *271*, 169-177. https://doi.org/https://doi.org/10.1016/j.jad.2020.03.081

Varnik, P. (2012). Suicide in the World. *International Journal of Environmental Research and Public Health*, *9*(3), 760-771. https://doi.org/10.3390/ijerph9030760

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 6000-6010.

Verma, I. M. (2014). Editorial Expression of Concern: Experimental evidence of massivescale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 201412469.

Wardhaugh, R. (2011). *An introduction to sociolinguistics* (Vol. 28). John Wiley & Sons.

Webb, S. (2018). Deep learning for biology. *Nature*, *554*(7693).

Wiegreffe, S., & Pinter, Y. (2019). Attention is not not Explanation. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 11-20.

Woods, W. C., Arizmendi, C., Gates, K. M., Stepp, S. D., Pilkonis, P. A., & Wright, A. G. C. (2020). Personalized models of psychopathology as contextualized dynamic processes: An example from individuals with borderline personality disorder. *Journal of Consulting and Clinical Psychology*, *88*(3), 240.

Wright, A. G. C., & Woods, W. C. (2020). Personalized Models of Psychopathology. *Annual Review of Clinical Psychology*, *16*(1), 49-74. https://doi.org/10.1146/annurev-clinpsy-102419-125032

Zaman, A. (2021). Combining Traditional and Non-Traditional Data Stream for Understanding Mental Health.

Zheng, L., Wang, O., Hao, S., Ye, C., Liu, M., Xia, M., Sabo, A. N., Markovic, L., Stearns, F., & Kanov, L. (2020). Development of an early-warning system for high-risk patients for suicide attempt using deep learning and electronic health records. *Translational psychiatry*, *10*(1), 1-10.

Zhou, P., Qi, Z., Zheng, S., Xu, J., Bao, H., & Xu, B. (2016). Text Classification Improved by Integrating Bidirectional LSTM with Two-dimensional Max Pooling. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 3485-3495.

Zirikly, A., Resnik, P., Uzuner, O., & Hollingshead, K. (2019). CLPsych 2019 shared task: Predicting the degree of suicide risk in Reddit posts. *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, 24-33.

Ziser, Y., & Reichart, R. (2017). Neural Structural Correspondence Learning for Domain Adaptation. *CoNLL 2017*, 400.

Ziser, Y., & Reichart, R. (2018). Pivot based language modeling for improved neural domain adaptation. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 1241-1251.