Stanford University
Human-Centered
Artificial Intelligence

FEBRUARY
2021

SPECIAL HEALTH CARE SERIES

AI-Enabled Depression Prediction Using Social Media

Johannes C. Eichstaedt

Natural language processing for mental health monitoring is an emerging use of AI that is poised to disrupt the landscape of the health care industry. As the profusion of social media platforms allows for a wider swathe of the population to share their thoughts and feelings with the world, users' posts and reactions extend the scope of medical screening methods for psychological disorders such as depression. Users are already being marketed to with sophistication based on these behaviors — why not leverage these technologies for public health?

To give some sense of scale for the unaddressed need, in the United States, between 7 and 26 percent of the population experiences depression each year, but only between 13 and 49 percent of those people receive treatment — **this means that in the US, there currently may be 30+ million people in need of but not receiving mental health care.** (Of note, these numbers are pre-COVID; early studies suggest that the prevalence of mental health conditions may have doubled after the first lockdowns). These high rates of underdiagnosis and undertreatment suggest that new screening methods like AI-enabled prediction are needed to identify and treat patients with depression.

In a recent article I co-authored in the *Proceedings of the National Academy of Sciences*, "Facebook language predicts depression in medical records," my team and I specify a set of protocols to **identify patients suffering from depression using only language from their Facebook posts**. These methods capitalize on significant advances in technology over the last decade and are capable of roughly matching the accuracy of

KEY TAKEAWAYS

- AI-enabled depression prediction is capable of matching the accuracy of traditional screening surveys but can be delivered to whole (consenting) populations.
- By examining social media language, our model can make a significant impact in recognizing the most widespread mental illnesses in the world.
- Policymakers and regulators must establish clearer guidelines about access to data, understand the consequences of using algorithms to change social media posts into protected health information, and consider how depression detection can be combined with digital treatments in a modern system of care.

traditional screening surveys. Using a system that relies on machine learning to cluster, count, and score each word, we find that language predictors of depression include emotional, interpersonal, and cognitive processes represented by words such as sadness, a preoccupation with the self or rumination, and expressions of loneliness and hostility.

In order for [depression assessment through social media] to become feasible as a scalable complement to existing screening and monitoring procedures, policymakers and regulators will need to ensure that patient privacy and confidentiality are kept at the forefront of consideration when these technology solutions are developed.

Depression assessment through social media represents a way to screen which does not require users to actively engage in a survey; it is unobtrusive for individuals who consent to be part of this modality. However, in order for this method to become feasible as a scalable complement to existing screening and monitoring procedures, policymakers and regulators will need to ensure that **patient privacy and confidentiality** are kept at the forefront of consideration when these technology solutions are developed. To that end, clearer guidelines and regulation are needed about access to data, who has access to the data, and the purpose for collecting this data. The application of machine learning to quasi-public social media posts can transform such data into protected health information and must be understood and treated as such — including with regards to questions of privacy and of the medical autonomy of patients.

If this advance in treatment is responsibly developed and is responsibly developed and introduced in a manner that integrates with existing systems of care and relationships with trusted providers, it has the potential to be a huge shift in public health.

Introduction: Language reveals health

Digital epidemiology as a field can be traced back about 10 years, when search engine data first allowed researchers to track flu trends in real-time, preceding hospital-reported indicators by about 2 weeks. The idea was soon extended to other “big data” sources and into the realm of medical psychology. By 2015, an important proof-of-concept had been published that linked the language people used on social media with health outcomes. I was part of a team that evaluated more than 100 million tweets from more than 1,300 counties

across the U.S., to determine that the **preponderance of negative tweets like those expressing anger or hostility reliably predicted rates of death from heart disease in a given county location**. Importantly, the people on Twitter are not the (generally, older) people dying from heart disease; instead, users on Twitter serve as "canaries in the coal mine" who let us see into the stress and hostility levels in communities, which in turn, affect those at highest risk. Surprisingly, using Twitter, we were able to predict county deaths from heart disease more accurately than comparable risk analyses made using government statistics for known risk factors (such as hypertension, diabetes, smoking, etc.).

Language contains a variety of signals that are not just how "positive" or "negative" the language sounds — an approach known as sentiment analysis. Based on a rich literature in psychology, we know, for example, that the use of "I," "me," and "mine" signals a preoccupation with the self that is almost always elevated in depressed individuals, and that robustly predicts negative emotionality. We know from both the psychological and the health literature that stronger emphasis on social connection and community ("we," "our" and family words) protects against mental illness and even heart disease (the "social buffering hypothesis"). We have found language indicating boredom to be unhealthy and excitement and interest to mark health and well-being — this dovetails with the observation that unemployment lowers well-being more than almost any other life event.

The language on social media feeds presents a particularly challenging set of data to parse because it is so filled with slang and emoticons. Thanks to breakthroughs in artificial intelligence and natural language processing made over the past decade, we are nevertheless able to extract meaningful patterns. With exponential growth in computing power, **it has become possible to process language using statistical pattern-recognition algorithms**—an approach that avoids the shortcoming of previous research programs that merely looked for keywords. The new generation of methods allows the important signal to emerge from the data. They cluster, count, and score words and phrases to learn psychological associations from scratch — letting the data tell its own story and ensuring that the researcher obtains prediction models that generalize robustly to new data. Parsing the language signal, researchers can draw appropriate connections between patient health and the ways they choose to express themselves — for example, for depression, we observe symptom clusters that span rumination, self-occupation, hostility, loneliness, and strained social relationships, self-doubt, and so forth. Perhaps surprisingly, large-scale, data-driven AI analyses have mostly pointed to processes and symptoms of depression proposed by the leading theories of depression, which increases our faith in these new A.I. methods (through "convergent validity" with the clinical literature).

Research with AI: Personality and depression

As Facebook, Twitter, and similar platforms have taken off during the past decade, the amount of data available for language analysis has expanded dramatically, offering psychologists a vast new window into the mental health of social media users.

We first explored these methods on the "standard model" of psychology: Big Five personality. In 2013, I was part of a team that published a study — since cited more than 1,000 times — in which we applied machine learning to 700 million words, phrases, and topics gleaned from the Facebook messages of 75,000 volunteers who had taken personality tests. We found that we can predict personality based on Facebook language and that the language associated with personality made a lot of sense: for example, extroverts use the word "party" more often, and introverts more often use "books." Conscientious individuals schedule both their work life and their downtime. Users who are open to experience are into "art" and "anime." More subtle signals also emerged: "apparently" is associated with neuroticism, team sports ("volleyball") with emotional stability. In recent work, we have applied these prediction models to Twitter to derive the personality composition of all counties in the US.

More recently, our analyses of language on social media have focused on mental health and on helping to identify depression in individual patients. By examining Facebook language data from a sample of consenting patients in a hospital setting, we built a method to predict if and when a person would receive a first diagnosis of depression in her electronic medical record by the hospital system, compared to a control group. **The performance of the algorithms roughly reached the minimal threshold for clinical usefulness**, comparable to that of validated depression scales. In light of how widely underdiagnosed depression is and how often the condition is missed by primary care doctors, this result is exciting because we can obtain these predictions without requiring time from doctors or patients — A.I. can simply analyze autobiographical texts of patients (like that on social media). In conjunction with a new generation of scalable digital treatments, this methodology makes possible a significant step forward towards widespread access to mental health care.

We further sought to investigate how far in advance we could predict future depression, since the earlier we can identify patients, the better the chances for treatment and for avoiding significant dysfunction in their lives. We found that social media-based prediction of future depression status may be possible as early as 3 months before the first documentation of depression in the medical record. Importantly, we believe that **AI-based analysis of social media should be part of a multi-step screening procedure**, in which the algorithms "raise red flags'' for some patients, who are then followed up with screening surveys and attention by a health care provider. There is a possible future in which these AI-systems complement the care provided by doctors: the algorithms can track depression severity over time and during treatment and provide a dashboard with specific symptoms and pain points to providers that they in turn can ask patients about. These methods aim to complement existing systems of care and extend their reach deeply into the digital worlds we already inhabit.

Policy Discussion

In our research, we always obtain permission to analyze participants' social media feeds, follow strict privacy guidelines, and are externally reviewed for compliance and ethical research conduct. But few social media users realize that giving access to their statuses (or even their likes) can supply a team of researchers (or a corporation) with a fairly fine-grained personality and mental health profile. Our findings raise important questions related to

patient privacy, informed consent, data protection, and data ownership. Analyses of how people use language on social media are based solely on statistical patterns, but they can be so revealing that intelligence agents, political candidates, and businesspeople ranging from marketers to insurance actuaries are just as interested in their application as scientists. In principle, these methods can be exploited without consent, as was the case in the Cambridge Analytica controversy — but the danger extends beyond personality to mental health status and membership in protected groups.

Developers need to address the possibility that the application of a medical algorithm to social media may change posts into protected health information, which in turn alters expectations around privacy and the right of patients to remain autonomous in their health care decisions.

Given the sensitivity of this content, clear guidelines are needed to govern access to these data, the people who have access, and those individuals' purpose. Developers need to address the possibility that the application of a medical algorithm to social media may change posts into protected health information, which in turn alters expectations around privacy and the right of patients to remain autonomous in their health care decisions. Similarly, those who interpret the data need to recognize that people may change what they write based on their perceptions of how that information might be observed and used.

To improve the accuracy of these digital screening methods, **developers and policymakers should look at how best to combine natural language from social media sites (and other data feeds like text messages analyzed within the phone) with a patient's location, and their physical activities and sleep patterns. However, they should also confront lingering ethical questions about how to use these predictions responsibly**. Carefully calibrated prediction systems that work for all genders and ethnicities (algorithmic fairness) and multi-step testing (with the addition of surveys and an expert-in-the-loop) will be a necessary part of the equation. These screening pipelines should integrate with emerging efforts to deliver mental health care that is both digital and scalable; modern "computerized Cognitive Behavioral Therapy," for example, is emerging as a possible pathway, particularly when delivered in a "blended" fashion that includes a health care provider in-the-loop to provide motivation and accountability. However, these potential treatments first require that patients have been identified; for this step, the new generation of AI-based methods seems to be one of the most promising pathways.

The more we realize the potential of our digital signals to give us insight into and improve our health and well-being, the more we can craft our mental health future in ways that are conscious, self-determined, ethical, and even lifesaving.

The original article, **"Facebook language predicts depression in medical records"** can be accessed at: https://www.pnas.org/content/pnas/115/44/11203.full.pdf

Stanford University's Institute for Human-Centered Artificial Intelligence (HAI), applies rigorous analysis and research to pressing policy questions on artificial intelligence. A pillar of HAI is to inform policymakers, industry leaders, and civil society by disseminating scholarship to a wide audience. HAI is a nonpartisan research institute, representing a range of voices. The views expressed in this policy brief reflect the views of the author. For further information, please contact **HAI-Policy@stanford.edu.**

Johannes C. Eichstaedt is a computational social scientist who is jointly appointed as an Assistant Professor at Stanford University's Department of Psychology and as the Shriram Faculty Fellow at the Institute for Human-Centered Artificial Intelligence (HAI). He is a co-founder of the World Well-Being Project.

Stanford University
Human-Centered
Artificial Intelligence

Stanford HAI: Cordura Hall, 210 Panama Street, Stanford, CA 94305-1234
T 650.725.4537 **F** 650.123.4567 **E** HAI-Policy@stanford.edu **hai.stanford.edu**