Mental Health Markers in Language and Brain Data: Potential Diagnostic Use and Privacy Concerns

Denisa Qori McDonald  
Drexel University  
Philadelphia, PA 19104, USA  
denisa.qori@drexel.edu

Girija Kaimal  
Drexel University  
Philadelphia, PA 19104, USA  
gk27@drexel.edu

Rachel Greenstadt  
Drexel University  
Philadelphia, PA 19104, USA  
rachel.a.greenstadt@drexel.edu

Erin T. Solovey  
Drexel University  
Philadelphia, PA 19104, USA  
erin.solovey@drexel.edu

License: The author(s) retain copyright, but ACM receives an exclusive publication license.

Abstract  
This paper discusses technology’s potential role in the inadvertent leaking of information related to mental health conditions, a particularly sensitive and legally protected part of one’s identity. Social media as well as emerging technologies such as brain-computer interfaces (BCIs) are changing the way that we interact with each other and the world. They also offer new windows into deeply personal and previously private aspects of our identity, and may leak this information. For example, with the vast amounts of public data on sites such as Twitter, we can now identify individuals even when they are using anonymous user accounts [35]. The internal state of an individual’s neural activity is, in many senses, the most private of personal data. Until recently it was impossible or highly inconvenient to gain access to this type of information. With these barriers lowering rapidly, it is critical to look carefully at the privacy implications.

Author Keywords  
mental health; design guidelines; privacy; discrimination; diagnosis; mental health detection; brain data; text data; fNIRS; stylometry; BCI.

Introduction  
Functional near-infrared spectroscopy (fNIRS) brain imaging is increasingly being used in mental health clinical settings [18, 50, 31, 27]. It has the potential for both contin-
ual lightweight monitoring and in diagnosis. At the same time, we have seen recent progress toward consumer-grade BCIs devices that are capable, affordable, safe, and portable, moving their use out of clinical settings and into new domains. There are proposals to leverage these devices’ ability to detect various cognitive states for applications such as authentication [10, 47], games, learning [46, 2, 20], and monitoring cognitive load in high-stress environments [4]. The fact that the same devices can be used for clinical purposes as well as entertainment and productivity purposes leads to a critical question: What personal health information might be leaked as people use BCIs in work or recreation? To facilitate the respectful usage of such technology, we must proactively study the feasibility of attacks on privacy, and explore potential mitigations.

Parallel to these developments, there has been increasing amounts of research in stylometry and sociolinguistics aimed at detecting and diagnosing mental health conditions from text, especially text on social media [11, 14, 12].

**Diagnostic & Side Channel Potential of Language**

*Stylometry* is the use of statistical analysis of written language, to uniquely identify a person. Features used may include the length of words, pairs of words, vocabulary usage, sentence structure, etc. Several resources give an overview of stylometry methods [51, 29]. Recent developments have led to robust classifiers using machine learning and other AI techniques [21, 49, 51]. This field has advanced sufficiently that its findings are routinely accepted as evidence in court [8].

*The study of how language is reflected in mental health is a relatively new, but active area of study.* Initial, pioneering work was done to quantify the psychometric properties of language use [38], and to build a program Linguistic Inquiry and Word Count to help in analysis. Language use has been shown to be a precursor of cognitive decline in studies of popular authors writing over time [52] and in the longitudinal Nuns study [42]. More recently, researchers have shown that it is possible to detect depression [13], PTSD [15], suicidal ideation [14], and many other mental health conditions via text published on social media platforms. In fact, for the past three years, there has been a workshop on Computational Linguistics and Clinical Psychology attached to the North American Association of Computational Linguistics (NAACL) conference devoted to the subject.

Stylometry is a wide field, and besides being recently used to explore the relationship between language and mental health, it has also been largely focused on profiling authorial traits [3, 16, 28, 33] such as gender, age, education, or even authorial personality [33, 40, 41]. For example, the detection of native language and language family from English text has been explored [45]. Some of this work has been used to predict sensitive information such as attrition in organizations [37]. *The ability to detect private information presents a significant privacy concern,* which will have implications as these methods are used for diagnosing mental health conditions such as Alzheimer’s, depression, psychopathy, even suicide risk.

**Diagnostic & Side Channel Potential of BCIs**

Several techniques measure the changing state of the brain: functional magnetic resonance imaging (fMRI), positron emission tomography, electroencephalography (EEG), and magnetoencephalography and functional near-infrared spectroscopy (fNIRS). EEG and fNIRS are the two main methods that have seen adoption outside of clinical settings due to their portability, relative low cost, and safety. EEG detects electrical impulses coming from the neurons firing in the brain, while fNIRS measures blood flow and blood oxygen changes re-
lated to the hemodynamic response, and is more similar to fMRI. The methods are complementary and can be measured simultaneously.

Research in mental health settings using fNIRS indicates that it could serve as a diagnostic tool for illness conditions related to affective disorders as well as disorders of self regulation. There is emerging evidence of biological correlates of mental status. For example, PTSD has been found to affect the speech production center of the brain and depression and schizophrenia have been associated with atypical activity [18, 26, 50]. fNIRS has been found to be related to specific brain pathways of reward perception in eating disorders, reduced pre-frontal cortex activation in depression and differential activation in schizophrenia [26, 31, 50]. These markers can be considered a brain signature for disorders.

Brain data can pose privacy threats as well. There has been considerable interest in its use as a biometric authenticator. Both fNIRS and EEG technology have been used as authentication metrics achieving high accuracies [36, 19, 39, 9]. Thorpe et al. [47] also researched the feasibility of using BCI for authentication, bringing up the ethical and privacy concerns of developing such a system. Venkatasubramanian et al. [53] show that certain physiological data is unique enough to be used to create an encryption key for sending patient data. Inherent in this is the fact that each person has a unique brain signature. Researchers have also explored side channel leakage of private information in BCI settings. For example, researchers have demonstrated the future potential for co-opting an EEG setup intended for one purpose, such as gaming, to extract unrelated, potentially private information [30].

Potential Scenarios in Mental Health

Research on reward pathways for eating disorders [32, 34] shows distinct patterns of participant gender and preferences which act as an individual signature that is objectively determinable using fNIRS. These markers can be considered a brain signature for disorders and potentially misused and abused by limited privacy protections. For example, the wireless fNIRS sensors available now could detect variations in reward perception leading to selective marketing to individual with a predisposition to addictive behavior and limitations in self regulation. Similarly, hemodynamic indicators of mood and affective disorders could be misused to discriminate and target patients. It might also be used to challenge diagnoses based on behavior or self report measures alone. For example, a patient’s request for disability or treatment might be challenged if a parallel fNIRS brain signature does not indicate the behavioral symptoms. This data might be collected as part of routine care without adequate informed consent. Patients need to be aware of the limits of privacy when participating in clinical procedures and the use of such data. Of particular note is the concern around the digital data generated from fNIRS. If not adequately protected, the digital data on hemodynamic responses could be manipulated and misrepresented by malicious individuals, as well as, for-profit ventures agencies seeking to make illegal profits including insurance companies, legal agencies and healthcare professionals. An additional issue with data generated by fNIRS is the archiving and secondary analysis of large databases as well as individual files for further analysis and research. Patients often might not know if and how their data might be used and what personal information might be shared as part of research or business and marketing efforts.

Users expose their language and brain data whenever they use brain-computer interfaces or enter textual information to online platforms, opening the possibility of the data being
used to infer private information about them. This might lead to unequal treatment, including in the workplace, in online forums, and less access to benefits that they would otherwise have access to.

**Toward Effective and Responsible Systems**

Given the potential for detecting mental health state that brain and text data can offer, when machine learning techniques are applied to them, it is important to design and build systems that leverage this capability that help the target population realize that they need help. At the same time, we should prevent mental health information from being inadvertently leaked and used by third parties putting subjects at a disadvantage.

We are currently investigating side channel leaks emerging in the use of brain-computer interfaces and social media, with a focus on mental illness. We combine our backgrounds in studying mental health including depression, disorders of self regulation and post traumatic stress disorders (PTSD) [55, 23, 25, 24, 54], with our work researching implicit brain-computer interfaces [43, 44, 48] and our experience investigating the privacy implications when using machine learning on complex, personal data [5, 6, 1, 7, 17, 22] to study privacy threats and develop appropriate mitigations. The questions described above will inform public debate on technologies that may leak private mental health information. This will enable the consideration of regulations early, before the technologies are widespread and their exploitation has already occurred. In addition, our findings in identifying mental health issues from everyday tasks could be integrated into future work in which these methods are used to inform the appropriate people (e.g. therapist, psychiatrist, or self) about a condition. When used appropriately, such a tool would be valuable in the assessment of mental health conditions and risks. The investigation of the potential disconnect between self-report and brain-report will have implications for brain-computer interfaces more generally. It will provide some early evidence about how much we can rely on brain data and how much people are able to manipulate it.

**References**


Information and Collective Privacy Behavior in a Large Social Network.


