# Integrating Online and Offline Data in Complex, Sensitive Problem Domains: Experiences from Mental Health

**Munmun De Choudhury**[†] and **Emre Kıcıman**[§]
[†]Georgia Institute of Technology, [§]Microsoft Research
[†]munmund@gatech.edu, [§]emrek@microsoft.com

## Abstract

A growing body of research in the ICWSM community and beyond has employed large-scale, unobtrusively gathered online data, primarily from social media sites, to model, understand, and rethink improving health and well-being. This short paper highlights the prior work of the authors in augmenting such online data driven approaches with offline information. To do so, the authors first present some of the challenges in utilizing online data alone in problems relating to the health domain. Then, we present three themes about how offline information may be harnessed, ranging from its use as a source of data, to obtaining theoretical explanations of computational models, and to improving the outcomes of online-data only models. Thereafter, we highlight some lessons learned from our work in doing so, in the domain of mental health. The paper concludes by situating offline information as an important resource that is critical to large-scale studies of health and well-being.

## 1 Introduction

Data-driven approaches have begun to make significant strides into impacting a variety of domains of broad societal significance (Boyd and Crawford 2012). A prime source of data that has particularly reaped widespread attention in the ICWSM community is online data, such as data stemming from people's social interactions on the web. Perhaps the biggest strength in the success of these online data in research stems from the fact that they now serve as an increasingly prominent platform of expression to many people globally (Perrin 2015); individuals can easily and quickly share their thoughts, ideas, opinions and events of interest with others. While such content sharing can be objective in nature, it can also reflect emotional states from the personal (e.g., loneliness, depression) to global scales (e.g., thoughts about a political candidate, musings about a newly released product or the global economy), and so on (Kaplan and Haenlein 2010). Consequently, online data has been employed to understand not only aspects of online interaction, social engagement, and information sharing, but also have been employed in domains as diverse as politics, economics, and health (John Walker 2014).



However, it is to be noted that problems in many domains like health are complex and sensitive, especially given that errors can be very costly. Not only do the problems involve deriving insights from the data, here online data, but it also means determining if those insights are practical and can be used to help relevant domain experts and stakeholders. Consequently, online data, alone, may provide only a limited perspective when the goal is to interpret and translate the methods and findings in real-world settings. This paper argues the need for augmenting online data with other forms of offline data, and situates the argument in the experiences and lessons learned by the authors in the mental health domain. We highlight how online and offline data complement each other, and also pose open challenges and opportunities to the ICWSM community.

## 2 The Limits of Online Data

Working with online data in a complex domain can be difficult for several reasons. Social data, in particular, is susceptible to a broad array of biases that challenge analysis (Olteanu et al. 2016). Here, we briefly discuss three of the broad issues we have come across when using online data to understand and address complex, sensitive domains.

**Construct validity.** First, when making measurements in online data, we must be concerned with the construct validity of our measures. That is, are we actually measuring what we think we are measuring? For example, if we are trying to measure mood from the language people use on social media, are the words they use reflective of the moods they are actually experiencing? While this may sometimes be the case, self-presentation bias, cultural norms, word ambiguities, and even song lyrics can complicate the association between people's experienced moods and their expression on social media. If not recognized and corrected, these false associations can entirely threaten the validity of our measurements and, through them, any conclusions we might wish to draw from the online data.

**Unobserved factors.** Second, when we are attempting to understand a phenomenon through its representation in online data, we must be aware that our observations may be significantly influenced by unobserved factors. When modeling people's behaviors and their reactions to an event or treatment within the online data, for example, we must take into account that people will also be affected by external cultural factors, social influence, seasonal dynamics, larger trends, and unobserved events (e.g., natural disasters). How these factors manifest can vary as well: each may vary across individuals in a data set, or alternatively affect all individuals simultaneously. These unobserved factors can confound our understanding of the situation, causing us to misunderstand the underlying mechanisms and draw the wrong conclusions about the severity of a situation or about the recommendations for action to improve a situation.

**Population biases.** Third, when learning about a complex domain from online data, we must have an understanding of the make-up of the populations we are studying. It is possible that our learnings from an online data source are only valid under certain situations or for a certain group of people. Because of the complexities of mental health, as well as many other domains, conclusions drawn based on a limited sub-population might be very different than conclusions drawn for another sub-population or for the population as a whole. If we wish to generalize what we are learning, we must have validation that the people and the specific situations we are studying through online data are representative of the broader phenomenon we care about.

These issues are not only concerns in online data, of course. Construct validity, unobserved factors, and population biases are challenges when analyzing any dataset, online or offline. We argue that these issues are particularly critical threats to the validity of online data analysis, however, because of the added distance that online data studies place between the scientist and the data generation process itself. We argue that studying online data to understand a complex phenomenon—and particularly to study interactions and developments occurring primarily offline—will almost necessarily require additional studies and validations that go beyond the online data itself to explicitly bridge the boundary between online and offline. For example, it is important to validate what our online data measures, and this validation must often come from sources outside our data (e.g., domain knowledge, external validation, active experiments). Similarly, surveying possible unobserved factors and understanding their potential effects, and characterizing the population represented within an online data set all require reaching beyond the online data.

## 3 Integrating Online and Offline Data

In this section, we discuss some key methods for augmenting online data with the help of offline information.

### Offline Data: Source of Gold Standard Information

One of the common places where researchers tend to leverage offline data in their social media data modeling and analyses lies in gathering gold standard information that can later be employed in supervised learning models. In the domain of mental health, this translates to compiling ground truth information about the true mental health states of individuals, communities, and populations, that is independently assessed outside of their online data.

In our prior work, we have extensively utilized this form of offline data. For instance, we used crowdsourcing, particularly through the Amazon Mechanical Turk platform, to collect (gold standard) assessments from several hundred Twitter users who reported that they have been diagnosed with clinical depression, using the CES-D (Center for Epidemiologic Studies Depression Scale) screening test (De Choudhury et al. 2013). Based on this cohort for whom we had offline assessments of depression, we developed several affective, behavioral, cognitive, linguistic, and domain-specific measures and use them to quantify an individual's social media behavior for a year in advance of their reported onset of depression, as assessed from their offline psychometric data. Then we leveraged these multiple types of signals from these measures to build a depression classifier, that can predict, ahead of onset time, whether an individual is vulnerable to depression. Our models show promise in predicting outcomes with an accuracy of 70% and precision of 0.74. Further, we evaluated this model by comparing it with gold standard offline statistics of prevalence of depression in the United States (De Choudhury, Counts, and Horvitz 2013). We found our social media index of depression compared well with these offline rates given by the Centers for Disease Control and Prevention. Similar approaches were used in other works from our team, including research that used Facebook data to predict risk of postpartum depression in new mothers (De Choudhury et al. 2014), and that employed (offline) clinical appraisals from clinical psychologists and psychiatrists to assess and curate quality of online data related to schizophrenia and psychosis (Birnbaum et al. 2017).

In a similar vein, in a different work (Chancellor et al. 2016), we employed feedback from clinical psychologists as gold standard information to infer mental health severity in pro-eating disorder posts on Instagram. Instead of getting expert annotations on posts directly, a method that does not scale well to large datasets, we obtained them on outcomes of topic models. This allowed us to scale our inference framework to a large corpus of Instagram posts, where we developed a semi-supervised approach to map the labels on the topics to posts from users.

### Offline Data: Interpreting Large-Scale Analysis

As noted above, online data, like that gathered from social media, can be complemented with offline data for interpreting the outcomes of an analysis or a computational model, or to contextualize the findings derived from online data in existing theory or theoretical frameworks.

In prior joint work of the authors (De Choudhury et al. 2016), the authors developed a causal inference framework (Pearl and others 2009) to assess the likelihood that an individual will transition to discussions of suicidal ideation, given a history of mental health discourse on social media. This framework was developed on a large dataset gathered from social media site Reddit. The output of the framework included words and phrases that indicated the likelihood of future suicidal ideation given their usage in a post. However, these linguistic cues did not allow us to examine how specific types of risk markers were associated with suicidal ideation, as illustrated in clinical psychology theories. To enable such comparison, we clustered these linguistic cues via spectral clustering to identify what themes led to increases or decreases in suicidal ideation. Then, we qualitatively interpreted these themes with the socio-cognitive model of suicide (Wenzel and Beck 2008), to understand what risk markers of suicide are manifested in social media, and to what extent the linguistic cue clusters align with what is known to exacerbate or alleviate the risk of suicidal ideation.

### Offline Data: Improving Computational Models

In this final subsection describing the co-utilization of online and offline data for health, we describe joint work of the authors in which offline insights were incorporated to revise and improve the outcomes of a computational framework. Such approaches can be a way to fill in the gaps left behind by use of online data alone, especially those gaps that are attributed to the limited 'view' on human behaviors and health provided by online data.

We briefly summarize such an approach from our prior work. Utilizing comments received on posts shared in Reddit mental health communities as a proxy for social support, in recent research (De Choudhury and Kıcıman 2017), we developed a human-machine hybrid statistical methodology that modeled and quantified the effects of the language of these comments in individuals who do and do not post on a suicide support community in Reddit. Applying stratified propensity score matching (Caliendo and Kopeinig 2008) in a iterative fashion, we first identified linguistic features of comments that showed significant effects. We obtained human assessments on the presence of suicidal ideation risk markers in posts associated with these features, for which the rater relied on their offline understanding and knowledge of the risk markers of suicide. Then we filtered the features that correspond to comparable subpopulations. Finally, we included these assessments in computing local average treatment effect, so as to assess the effects of specific linguistic features of comments in future risk to suicidal ideation.

## 4 Lessons Learned, Outstanding Challenges

In conducting the research discussed above, wherein we employed offline information in various ways to augment the computational models that utilize online social media data, we encountered a number of challenges that need careful attention and consideration. We present these lessons as calls to action for the broader research community invested in this line of research.

**Quality of Offline Data, Replicability of Offline Data Curation Approaches.** Just as collecting and curating online data is critical to the success of computational approaches that seek to understand and infer health, it is critical to also ensure the use of quality and reliable offline data. For instance, if the offline data is a source of gold standard information, clear guidelines for annotations that are grounded in the literature, assessment of quality and reliability, as well as scalable and objective means of collating such data that can be replicated to other contexts and problems are important considerations. These considerations can be factored in via sound statistical techniques of reliability and quality assessment, qualitative examination, as well as via study and experimental designs that enable transparency and replication of the curation and gathering process of offline data.

**Collaboration with Domain Experts.** In our prior research, offline information came from domain experts, who provided context and grounding to the analysis provided by the online data. While there can be many possible sources and means to obtain offline data, assessment of the expertise and credibility of that source is paramount. We emphasize that in most problems, ongoing collaborations between computational researchers and domain experts such as from the fields of social sciences, health, and medicine can help better ground the online data analysis and foster more robust findings and outcomes by the incorporation of independent but complementary offline, theoretical/practical insights.

**Theoretical Investigation.** A lesson that we gleaned from our prior research is the significance of theory, that is typically built on offline insights and offline studies, in augmenting online data-driven studies of health. As discussed above, theory can help contextualize the empirical observations gathered from online data. However, many important questions remain, as to how and what need to be the best practices to utilize theory as a source of offline insights. How can theoretical frameworks of health challenges translate to online data? How can we operationalize theoretical offline insights and observations in the online domain? What should researchers do when this transformation or operationalization is not possible? What happens when outcomes of models utilizing online data and that given in the theory do not agree or align? To navigate these challenges, we suggest more work in the area of merging online and offline data, and stronger interdisciplinary collaborations.

**Translating Insights into Interventions.** Finally, the underlying impetus for investigating problems of societal significance is, of course, the desire to help people improve their outcomes, whether through early identification of people at risk, better personalization of treatments, or discovery of new treatment strategies. Bridging the gap between insights derived from online data and real-world (offline) action will require combining continuous online data collection and analysis; interventions (whether online or offline); and simultaneous offline observations to provide strong validations of benefits. The challenges posed in moving from analysis to intervention in online platforms are exacerbated in sensitive domains—e.g., how to get informed consent from very large populations, or how to ensure interventions avoid real-world harms while respecting privacy of individuals online. It will be a significant challenge to develop new protocols that safely translate insights from observational studies to active experimentation and then to large-scale deployments, while simultaneously respecting principles of individual autonomy, minimizing risk of harm, and ensuring benefits and risks are justly distributed across participants.

## 5 Conclusion

We reviewed some of our prior work wherein we utilized offline information, in the form of data, insights, theory, and expert feedback to augment online data based computational models for health. The discussion hopes to initiate deeper conversations between interdisciplinary scientists as to how both these types of data can be helpful in tackling and addressing outstanding challenges pertaining to individual and population-level health and well-being.

## References


Birnbaum, M. L.; Ernala, S. K.; Rizvi, A. F.; De Choudhury, M.; and Kane, J. M. 2017. A collaborative approach to identifying social media markers of schizophrenia by employing machine learning and clinical appraisals. *Journal of medical Internet research* 19(8).

Boyd, D., and Crawford, K. 2012. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society* 15(5):662–679.

Caliendo, M., and Kopeinig, S. 2008. Some practical guidance for the implementation of propensity score matching. *Journal of economic surveys* 22(1):31–72.

Chancellor, S.; Lin, Z. J.; Goodman, E.; Zerwas, S.; and De Choudhury, M. 2016. Quantifying and predicting mental illness severity in online pro-eating disorder communities. In *Proceedings of the 19th ACM conference on Computer supported cooperative work & social computing*, 626–638.

De Choudhury, M., and Kıcıman, E. 2017. The language of social support in social media and its effect on suicidal ideation risk. In *ICWSM*, 32–41.

De Choudhury, M.; Gamon, M.; Counts, S.; and Horvitz, E. 2013. Predicting depression via social media. In *AAAI Conference on Weblogs and Social Media*.

De Choudhury, M.; Counts, S.; Horvitz, E.; and Hoff, A. 2014. Characterizing and predicting postpartum depression from facebook data. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM.

De Choudhury, M.; Kıcıman, E.; Dredze, M.; Coppersmith, G.; and Kumar, M. 2016. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2098–2110. ACM.

De Choudhury, M.; Counts, S.; and Horvitz, E. 2013. Social media as a measurement tool of depression in populations. In *Proceedings of the 5th Annual ACM Web Science Conference*, 47–56. ACM.

John Walker, S. 2014. Big data: A revolution that will transform how we live, work, and think.

Kaplan, A. M., and Haenlein, M. 2010. Users of the world, unite! the challenges and opportunities of social media. *Business horizons* 53(1):59–68.

Olteanu, A.; Castillo, C.; Diaz, F.; and Kıcıman, E. 2016. Social data: Biases, methodological pitfalls, and ethical boundaries. Available at SSRN: https://ssrn.com/abstract=2886526 or http://dx.doi.org/10.2139/ssrn.2886526.

Pearl, J., et al. 2009. Causal inference in statistics: An overview. *Statistics surveys* 3:96–146.

Perrin, A. 2015. Social media usage: 2005-2015.

Wenzel, A., and Beck, A. T. 2008. A cognitive model of suicidal behavior: Theory and treatment. *Applied and preventive psychology* 12(4):189–201.