TBM

ORIGINAL RESEARCH

A computational study of mental health awareness campaigns on social media

Koustuv Saha,[1,*] John Torous,[2,*] Sindhu Kiranmai Ernala,[1] Conor Rizuto,[2] Amanda Stafford,[2] Munmun De Choudhury[1]

[1]School of Interactive Computing, College of Computing, Georgia Institute of Technology, Atlanta, GA 30332, USA

[2]Division of Digital Psychiatry, Department of Psychiatry, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA 02115, USA

*Co-first authors.

Correspondence to: J Torous, jtorous@bidmc.harvard.edu

Cite this as: *TBM* 2019;XX:XX–XX
doi: 10.1093/tbm/ibz028

ABSTRACT

As public discourse continues to progress online, it is important for mental health advocates, public health officials, and other curious parties and stakeholders, ranging from researchers, to those affected by the issue, to be aware of the advancing new mediums in which the public can share content ranging from useful resources and self-help tips to personal struggles with respect to both illness and its stigmatization. A better understanding of this new public discourse on mental health, often framed as *social media campaigns*, can help perpetuate the allocation of sparse mental health resources, the need for educational awareness, and the usefulness of community, with an opportunity to reach those seeking help at the right moment. The objective of this study was to understand the nature of and engagement around mental health content shared on mental health campaigns, specifically #MyTipsForMentalHealth on Twitter around World Mental Health Awareness Day in 2017. We collected 14,217 Twitter posts from 10,805 unique users between September and October 2017 that contained the hashtag #MyTipsForMentalHealth. With the involvement of domain experts, we hand-labeled 700 posts and categorized them as (a) Fact, (b) Stigmatizing, (c) Inspirational, (d) Medical/Clinical Tip, (e) Resource Related, (f) Lifestyle or Social Tip or Personal View, and (g) Off Topic. After creating a "seed" machine learning classifier, we used both unsupervised and semi supervised methods to classify posts into the various expert identified topical categories. We also performed a content analysis to understand how information on different topics spread through social networks. Our support vector machine classification algorithm achieved a mean cross-validation accuracy of 0.81 and accuracy of 0.64 on unseen data. We found that inspirational Twitter posts were the most spread with a mean of 4.17 retweets, and stigmatizing content was second with a mean of 3.66 retweets. Classification of social media–related mental health interactions offers valuable insights on public sentiment as well as a window into the evolving world of online self-help and the varied resources within. Our results suggest an important role for social media–based peer support to not only guide information seekers to useful content and local resources but also illuminate the socially-insular aspects of stigmatization. However, our results also reflect the challenges of quantifying the heterogeneity of mental health content on social media and the need for novel machine learning methods customized to the challenges of the field.

Keywords

Social media, Twitter, Mental health, Machine learning, Public health

Implications

Practice: Online social media campaigns offer popular forums for the public to share their thoughts, advice, and information on resources about mental health.

Policy: Automatically classifying Twitter-based mental health content remains a challenge that limits the use of this data to inform policy.

Research: New collaborations between patients, clinicians, and data scientists are necessary to better understand the sentiment and spread of social media–based mental health campaigns.

INTRODUCTION

With depression a leading cause of global disability [1] and the burden of mental health conditions projected to continue to rise [2], there is an urgent need for new solutions and tools for mental health [3]. Considering that 300 million people worldwide have depression, 60 million have bipolar disorder, and 23 million have schizophrenia [1]—it is clear that any potential solution must leverage scalable technology able to reach the millions in need and billions at risk.

Social media are described as Internet-based applications, which allow people to share opinions [4]. Social media are considered among the mass media communication channels—together with newspapers, magazines, billboards, radio, television, and Internet—but they are distinct in that they enable people to be actively involved in the communication process and stay connected with other [4]. It has been well documented that social media constitutes an immensely powerful source of social influence [5], with an ability to help individuals frame opinions on topics they care about, or to alter attitudes and perceptions around events and issues [6]. With Facebook harboring over a billion users, over 2.5 billion active users of social media today, and with expanding penetration in high-, medium-, and low-income countries [7], these technologies can be

Published online: XX XXXX 2019

construed to be a key element of any technology-facilitated mental health solution [8].

One of the biggest strengths that social media provides revolves around its ability to reach large populations quickly, inexpensively, and with low effort [4]. Consequently, in recent years, a more developed use of social media for improving mental health has been through raising awareness, conducting outreach, and forecasting trends [9–13]. Social media–based mental health campaigns such as the Bell Let's Talk effort in Canada have been temporally associated with an increased rate of mental health visits among youth [14]. In fact, a review of such social media mental health campaigns has also found that these platforms hold promise in changing user behaviors, and that they are highly effective in recruiting participants and motivating them to take small, concrete actions [15]. It has also been demonstrated that there is room in social media for targeted, inexpensive, small-scale projects, as well as large, well-funded, mass-reach marketing blitzes [16]. In other relevant work, the social media platform Twitter has been used to help forecast the acuity of health emergencies in real time such as the 2013 Boston Marathon bombings [17] as well as offering population level of data on suicide risk factors [18]. Researchers have also developed methods to predict individual-centric mental illness diagnoses based on Twitter posting [19], identify spikes in use related to mental health [20], study population-level mental health awareness on Facebook [9], and classify posts based on lexical information and emotions [21]. However, less is known about the actual content, types of messages, and information being shared on mental health through Twitter, particularly surrounding raising or improving awareness and understanding public opinion and sentiment.

Relating to the potential of social media as a platform to improve mental health literacy, recent studies suggest that people use Twitter to discuss mental health to build community, raise awareness, have a safe place to express themselves and discuss personal struggles, serve as a coping mechanism, get peer advice and help, and appropriate it as a tool toward empowerment [22]. In fact, it has been argued that the diverse communities on social media help to make mental illnesses, which are often invisible to friends and family, visible through postings, photos, and videos [16, 23], and thereby can support altering perceptions of stigma [24]. But there is also evidence that mental health stigma has migrated online and some Twitter posts are inappropriate and condescending toward those with mental health conditions [25]. Although it is clear that those with mental illnesses use Twitter to talk about their lived experiences [26], it remains unclear if mental health tagged content is primarily being shared to offer peer advice on treatment, inspiration and hope, resources, facts, or outright inaccurate information. Understanding the types of content being shared on Twitter, centered on mental health outreach is critical to a number of possible mental health applications and uses. These may include, assessing its public health impact as well as public attitudes, identifying potential for interventions such as fighting stigma, influencing public health policy decisions, tailoring mental health literacy efforts, and helping scale positive uses such as peer or technology-driven or assisted support or information on meaningful resources or promoting positive behavior change [27].

Consequently, in recent years, a number of mental health campaigns have surfaced that have either primarily evolved via social media such as Twitter or used social media as a channel of communication and a mechanism to reach wide, national and global populations. Many social media platforms and activists have spearheaded these initiatives. For instance, in 2017, Instagram launched the #HereForYou campaign to help users find resources and support online and offline for how to get help with preventing and recovering from mental illnesses.

To realize the potential of Twitter data to inform public health campaigns, this article focuses on one such social media–based mental health campaign that gained significant traction on Twitter around the World Mental Health Awareness Day in 2017. The research seeks to identify the nature of this content shared on Twitter, and how different individuals engaged with it over a period of time. To do so, the article leverages expert assessments of mental health content to provide automated, robust, and scalable machine learning and statistical methods for content categorization and for understanding engagement. This way, our novel approach goes beyond a reliance on surveys and traditional media anchoring effects to understand public attitudes surrounding mental health—a gap noted in prior literature [28]. In addition, this work innovates over state of the art techniques to understand social media mental health content (e.g., see ref. [24]) that largely focuses on qualitative characterization and categorization of mental health awareness content on social media, requiring extensive effort- and time-consuming feedback from domain experts.

METHODS

Toward our goal of understanding the nature of and engagement around mental health content shared on mental health campaigns, we focus on a recent Twitter campaign, which started close to the World Mental Health Awareness Day (October 10, 2017). The campaign was spearheaded by the hashtag, #MyTipsForMentalHealth. It was a grassroots campaign that trended globally with over 30,000 Twitter posts in September 2017 [29, 30]. Our rationale behind choosing this specific campaign stems for two reasons: (a) Unlike other conditions, there have been very few mental health-specific campaigns on social media; and (b) this campaign had emerged as one of the largest ones in recent history, and the context of its inception around the World Mental Health Awareness Day allowed greater reachability to and awareness among diverse audiences. With the support of the creator of this campaign as an author, this social media campaign represented the intersection of clinical interest, mental health advocacy, and data and computer science.

Surrounding this specific campaign, our technical approach to address the proposed research goal involves the following steps: (a) collecting relevant social media data; (b) generating thematic annotations on a sample of these collected Twitter posts; (c) developing a machine learning framework to leverage the expert annotations and automatically infer the topic of a mental awareness campaign posts; and finally (d) building an analytical framework that appropriates the outcomes of the machine learning framework to explore the characteristics of content shared around mental health awareness as well as the general Twitter audience's engagement around this content.

Twitter data collection

We identified two trending hashtags on Twitter via which this mental health campaign spread: *#mentalhealthday* and *#mytipsformentalhealth* [29, 30]. Using these hashtags as search queries, we programmatically (and automatically without active human intervention) collected a large dataset of relevant Twitter posts. Specifically, our data collection approach used a web crawler-based Twitter Application Programming Interface (API): GetOldTweets. Our data collection spanned between September 01, 2017 and November 05, 2017, which collected *all* of the tweets associated with this hashtag. Our motivation behind using this specific API was that it provided all Twitter posts on a given search query (here a hashtag). Hence, our dataset did not suffer from any biases resulting from specific sampling strategies.

We obtained 14,217 Twitter posts that were shared by 10,805 unique users at an average of 1.32 posts per user. Corresponding to each post, we additionally obtained their engagement metrics in terms of the number of retweets (a signal of reshare) and favorites (a signal of endorsement). Note that, these two measures have been situated in prior literature as reliable indicators of content engagement on Twitter [31]. Table 1 reports the descriptive statistics of our dataset, Fig. 1a shows the daily volume of these posts in this period—we notice one sharp rise in the number of Twitter posts (with these hashtags) in the last week of September, and another one exactly on the World Mental Health Day (October 10) [32]. Figure 1b shows the engagement distribution of these Twitter posts. We also extracted the most frequently occurring hashtags in these posts (Fig. 1c) and find that these campaigns frequently use other hashtags that are contextually related to the campaign, such as *#mentalhealth* (335 posts), *#mondaymotivation* (199 posts), *#depression* (72 posts), and *#anxiety* (49 posts).

Further, we collected a variety of user attributes, such as the number of posts they had shared on Twitter, followers, and followees—this was essential to understand who engages with these content. For this, we could obtain the user attributes of 10,680 users who had public non-deleted accounts as of November 6, 2017. We found that these users demonstrated varied social media usage patterns, as indicated in terms of their number of posts ranging between 1 and 1,140,627 (Median = 5605.50, stdev. = 3,6118.5), and their followers to followee ratio, (which roughly estimates the popularity of a user is on Twitter) ranging between 0.01 and 54,802.47 (median = 0.97, stdev. = 723.3).

To further understand engagement, for each post in the dataset, we also obtained the number of retweets and favorites received by that post using the official Twitter API and via a technique developed in-house, which used parsing the HTML content of the links corresponding to each Twitter post. Overall, the 14,217 posts collected using the awareness hashtags had a total number of 48,223 retweets (mean = 3.39, stdev. = 29.3; Table 1), whereas they were favorited a total of 145,682 times (mean = 10.25, stdev. = 87.4; Table 1).

Expert labeling

Using a mini-modified Delphi process [33], the following eight categories for classification of Twitter posts in our dataset were identified as: (a) Fact (F), (b) Stigmatizing (S), (c) Inspirational (I), (d) Medical/ Clinical Tip (M), (e) Resource Related (R), (f) Lifestyle or Social Tip (LS), (g) Personal View (PV), and (h) Off Topic (OT). Randomly sampled Twitter posts (700 in all) were individually hand-labeled by a board-certified psychiatrist (coauthor Torous) and master's level psychiatry research

Table 1 | Descriptive statistics of our Mental Health Awareness Campaign dataset

Metric	Value	Metric	Value	Metric	Value
No. of posts	14,217	Number of retweets	48,223	Number of favorites	145,682
No. of users	10,805	Median retweets	3.39	Median favorites	10.25
Avg. posts per user	1.32	Stdev. retweets	29.28	Stdev. favorites	87.40

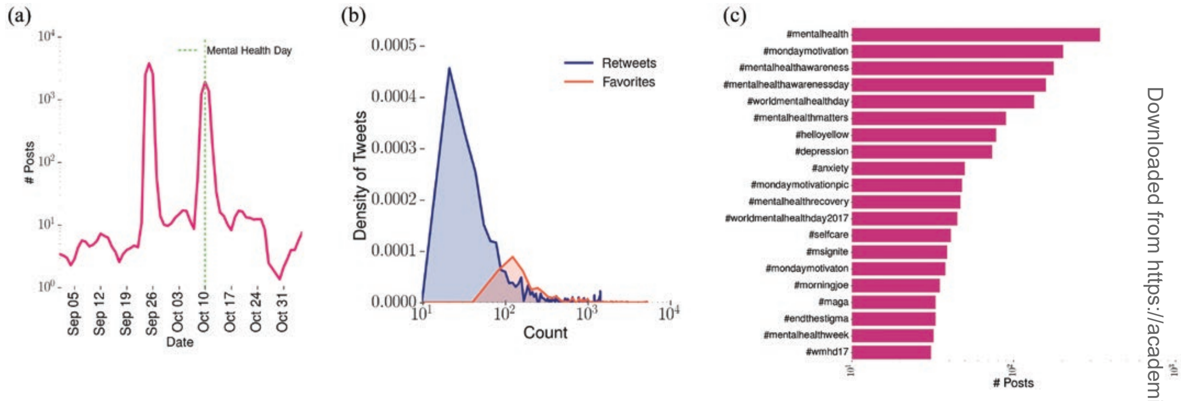

Fig. 1 | (a) Daily occurrence of MHAC posts; (b) distribution of retweets and favorites; and (c) top 20 most frequently used hashtags in the mental health awareness campaign (MHAC) posts.

assistant (coauthor Rizuto) who are familiar with Twitter and social media. Any disagreements (n = 42) were discussed by co-authors Torous and Rizuto together in person until 100% consensus was obtained. We provide example Twitter posts labeled corresponding to each of these categories in section 1 of Supplementary Material document.

Our largest labeled topic was LS with which occurred in 31% (219) of the posts. However, as we found that only four posts belonged to the Fact category, we did not consider this category as a distinct category for our downstream tasks. In addition, for better clarity and demarcation across the categories, we merged PV and LS into a single category of PL (Personal, Lifestyle and Social Tip). Our final set comprised six categories of Twitter posts: OT, S, I, M, R, and PL.

Machine learning approach to infer topics

Once we had expert-labeled a sample of 700 posts, our next objective was to use this sample to automatically infer the topic of all (of the remaining 13,517) Twitter posts in our data—this would give us a sense of the range of issues that surfaced in the specific Twitter mental health campaign of our focus. To do so, this work built a two-phase multi-class machine learning classifier, as described next. In machine learning, classification is the problem of identifying to which of a set of categories (subpopulations) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known.

Seed classifier (C_0)

From the 696 manually labeled Twitter posts (after discarding posts labeled as F), we held out 140 posts as our test dataset and used the remaining 556 posts as our training data for building a first machine learning classifier. To build this classification model (C_0), we used a variety of features that consider the structural, linguistic, and lexical aspects of Twitter posts, drawing on prior work on social media and mental health [25]. A description of these features is given in section 2 of Supplementary Material. Using a total of 634 features, we trained many classifiers with algorithms such as Random Forest, Support Vector Machines (SVMs), and Logistic Regression, as is standard practice for multi-class classification. We used k-fold (k = 5) cross-validation [34] for parameter tuning, and tested our seed classification models on the held-out test dataset. A short technical description of these classification models is given in section 2 of Supplementary Material.

Semi-supervised classifier (C)

The second phase of our machine learning approach seeks to improve on our ability to categorize the Twitter posts, by leveraging the seed classifier described earlier. This is essential due to the fact that in many real-world scenarios involving automatic categorization of content, like ours, and because labeled data are both expensive and time-consuming to gather as well as scarce (e.g., needs expert involvement and manual labor), although unlabeled data are comparatively huge and easier to gather. To build robust classification models that can generalize across datasets, settings, and mental health campaigns, we use semi-supervised learning [35–37] in this second phase, that is able to leverage both labeled and unlabeled data in unison, thereby is to cover a better diversity of training examples [38].

Accordingly, from the 13,517 unlabeled posts, we first obtained a random sample of 8,110 posts (i.e., 60% unlabeled dataset), from which we then found those posts that were similar to our labeled examples [39], on the basis of their social media structural, linguistic structure and affect, psycholinguistics, and n-gram features (section 2 of Supplementary Material elaborates on the features). In particular, we obtained these similar examples using k-Nearest Neighbor technique to significantly expand our limited sized hand-labeled training dataset of 696 posts with another 1,321 training examples (see section 2 of Supplementary Material for more detail). Our net expanded training dataset thus comprised 2,017 posts. We used the same set of features and repeated building several classifiers (C) according to the same methods as used for C_0 (ref. section 2 of Supplementary Material), and tuned their parameters with a *k*-fold ($k = 5$) cross-validation as described earlier.

Analytic approach of machine-labeled content and their engagement data

To understand in what ways the categories of Twitter mental health campaign content, identified earlier, differ in their content, we used an unsupervised language modeling technique [40]. In section 3 of Supplementary Material, we provide details of this approach. Finally, we studied the engagement received per each category of posts to understand the reach and impact of different topical categories. For each of the categories, we examined the probability distribution of retweets and favorites received by posts belong to the category.

RESULTS

Assessing the performance of the machine learning classifiers

We first report the *k*-fold ($k = 5$) cross-validation and test accuracy metrics of our preliminary seed classifiers (C_0). We find that the best model (based on mean and standard deviation of accuracy) is an SVM classifier that achieves a mean cross-validation accuracy of 0.50 (stdev. = 0.01) and a test accuracy of 0.52. This test accuracy is only slightly better than a baseline accuracy of 0.44, which is obtained by assigning all posts the label of the largest sized category. Detailed information about the cross-validation accuracy of seed classifier (C_0) is given in section 2 of Supplementary Material.

Next, we examine the accuracy of our semi-supervised classification approach. Our best model (C) used an SVM algorithm to achieve a mean cross-validation accuracy of 0.81 (stdev. = 0.08). This model predicts with a test accuracy of 0.64 on the unseen held-out dataset, showing an improvement by 25% from the seed classifier C_0. We refer the reader to section 2 of Supplementary Material for additional information about the performance of these classifiers.

Considering the top significant features of the classifier C, we find that certain hashtags such as *#mondaymotivation* and *#mentalhealthawareness* and the presence of a URL are significant contributors. This observation pertains to the very definition of our categories, where three categories are related to resources or tips. The other important features are predominantly *n*-grams, among which we find many terms that are related to health and mental health (*anxiety depression, stress, physical, etc.*).

We note that the Linguistic Inquiry and Word Count (LIWC) category of second person pronoun and imperatives (e.g., *look, ask, don't listen, let tell*) plays a significant role in our classifier. Prior research has found that the presence of second person pronouns in conjunction with interactive verbs *(ask, tell, listen, etc.)* is typically associated with social processes and social interaction [41], aspects key to raising awareness and dissemination of social media information, and therefore likely to surface in many categories in our dataset. The top features that belong to parts of speech (POS) sequences also convey that our classifier is able to capture differences in particular linguistic structures of expression. For instance, the sequence of "VB IN PRP" (verb–preposition–personal pronoun) captures interactive and imperative opinions (e.g., "*Know that you are not alone,*" "*don't worry about me,*" "*never forget that you are important":* the underlined segment denotes the particular parts-of-speech sequence). We refer the readers to section 2 of Supplementary Material for an extended list of these top significant features.

Analysis of machine-labeled content

After using this well-performing semi-supervised classifier C to automatically label all of the remaining 12,196 unannotated Twitter posts, we present an analysis of the different topic categories characterizing these posts and the ways in which the Twitter community engages with this content. Figure 2 presents the distribution of the posts per category in our data, and Fig. 3 and Table 2 present the distribution of the dataset's retweets and favorites per category.

Content analysis

We find that the Personal/Social Tip (PS) category shows the greatest occurrence, occurring in 44% posts (5,997 posts). This category includes personal-, social-, and lifestyle-oriented views and tips, and over 96% posts in this category are associated with the hashtag *#mytipsformentalhealth*. On the basis of the unsupervised language modeling approach introduced in Methods section (also see section 3 of Supplementary Material), we found this category to include phrases of advice and guidance relating to navigating one's mental health, such as "dont let," "think positive," "practice gratitude," "avoid toxic people," and "avoid news". Next, the Inspirational (I) category occurs in over 28% posts (3,822 posts). We find that the posts assigned to this category express the importance and positives of mental health and aim to encourage the individuals to actively seek mental health care when in need, as evident from phrases such as "okay ask help" and "health matters". Here, we also find phrases that hint a motivational tone such as "keep going" and "youre loved". The Resources-related (R) category occurs in almost 15% posts (1,937), and over 82%

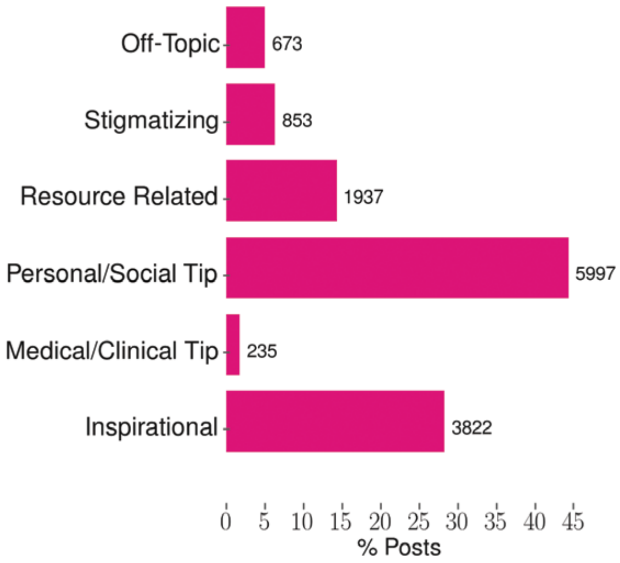

Fig. 2 | Posts per category labeled in our entire dataset of 13,517 posts.

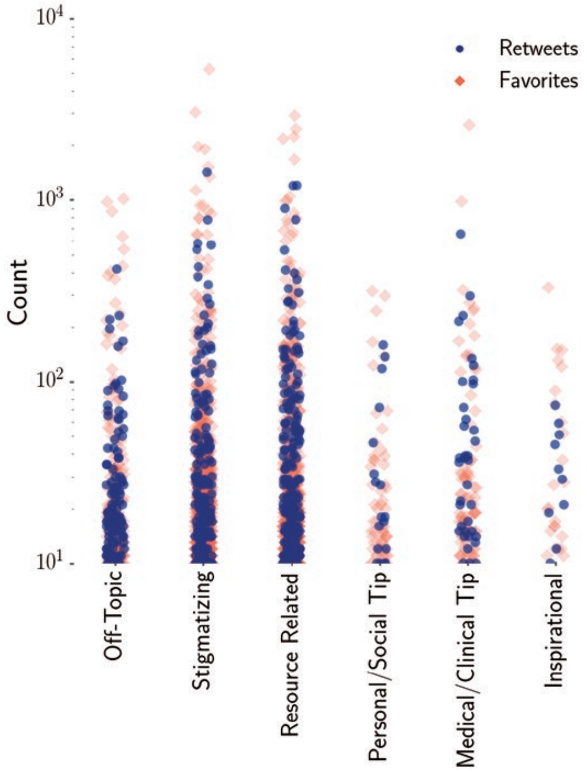

Fig. 3 | Distribution of retweets and favorites across the topical categories.

of these posts contain a URL. The top phrases in this category also demonstrate that these posts point resources such as reports, surveys, or information about mental health shows on Television and other media. Among the other relevant categories, we find that the Stigmatizing category that occurs in 6% posts advice the audience to stay away from particular media sources and also express political content. The Medical/Clinical Tip category occurs the least and this kind of content can be found in only about 2% posts. The top phrases in this category include keywords related to clinical professionals, medicines, health conditions, or treatment and therapy.

Engagement analysis

Across the entire corpus of 13,517 Twitter posts labeled by the classifier together with the 700 hand-labeled posts, there were 48,223 retweets (mean = 2.36, stdev. = 23.3) and 145,682 favorites

Table 2 | Engagement received in retweets and favorites per topic category

Category	#Retweets	Mean #RTs	#Favorites	Mean #Fav.
Stigmatizing (S)	3,350	3.66	9,199	10.05
Inspirational (I)	16,516	4.17	57,469	14.51
Medical/Clinical Tip (M)	560	2.20	1934	7.58
Resource Related (R)	6,378	3.17	15,544	7.72
Personal/Social Tip (PS)	20,127	3.17	57,691	9.09
Off Topic (OT)	1292	1.80	3,845	5.35

(mean = 6.68, stdev. = 67.1). Applying the data collection steps from earlier, we obtained a set of 14,642 unique users who retweeted the posts and 45,744 unique users who liked or favorited the posts, leading to a total of 60,386 users who engaged and responded to the 13,517 posts in all. Next, incorporating the machine labels from the classifier, the engagement received per category of posts is presented in Table 2. Kruskal–Wallis tests suggest statistical significance in both the number of retweets and favorites across the categories with $p < 0.001$, with corrected H-statistic values of 216.74 (for retweets) and 534.31 (for favorites).

We observe that engagement received (per post) through favorites is higher than that received by retweets, across all categories of posts. Examining the engagement received per each category, stigmatizing and inspirational posts can be noted to receive the highest engagement from an audience on Twitter. On the other hand, off-topic posts receive the lowest engagement both through retweets and favorites.

DISCUSSION

Our results underscore the potential of social media platforms such as Twitter to quantify the content, spread, and reach of mental health outreach efforts. A large body of literature has examined the potential and power of social media to initiate, drive, and engage publics around campaigns targeting various social, political, and health issues [42, 43], including instances when social media was the centerpiece of a campaign with print and television media used primarily to support the social media focus [43]. Our work contributes to and extends that line of research, although it does not investigate how campaigns that unfold over social media compared with those that are primarily concentrated on other mass media channels. As the communications landscape gets denser, more complex, and more participatory, the networked population is gaining greater access to information, more opportunities to engage in public speech, and an enhanced ability to undertake collective action [44]. These increased freedoms can and have helped loosely coordinated public's demand and drive change through increased reach [45]. In fact, today millions of messages can be shared easily to coordinate a diverse, rapid response that can lead to changes in attitudes, perspectives, and even regulations around important societal concerns, including health [15, 46]. Therefore, as we observe in our work and as has been argued for other types of health concerns [23], social media campaigns on mental health topics have an immense role to play to raise awareness, reduce stigma, support individuals living through these chronic conditions, or inexpensively but meaningfully engage otherwise passive bystanders.

Essentially, our work can pave the way for further research on unraveling the role of social media and other similar web-based channels in altering the public discourse on mental illnesses. In this regard, as has also been noted by Korda and Itani [46] and Freeman et al. [15], although campaign impact evaluation measures are available, new measures are needed to evaluate the effectiveness of social media campaigns, such as the one analyzed in this article. We are in agreement with other scholars [15] who noted that there is a need to incorporate outcomes, research, and theory in designing social media–based mental health promotion programs that can build on empirical observations of campaigns like the one this article presents.

Methodologically speaking, a key novelty of our work is that it reduces the need to incorporate extensive expert coded qualitative data for the purpose—an approach widely prevalent this far [22, 24] —by making principled use of machine learning and access to vast amounts of unlabeled social media data. Nevertheless, using semi-supervised machine learning methods, the 0.64 mean accuracy of our methods reflects the challenges of quantifying the heterogeneity of mental health content. But these results also underscore the potential to automate classification in a manner suitable for real-time population-level insights.

Our results raise questions regarding what level of classification accuracy for social media content is necessary to inform public health campaigns and measure their impact. There is currently no gold standard, and we propose that our results may set at least a starting point for future efforts to improve upon. The difficulty in building a fully automated classifier may in part be because mental health itself is a very heterogeneous term representing a series of further heterogeneous conditions. Although mental health conditions are brain-based illnesses [47], societal and cultural factors shape how individuals communicate and understand their experience of a mental illness [48]. Thus, the biological heterogeneity combined with societal and cultural variation offers a plausible reason why semi-supervised, versus unsupervised, methods may be of more value and produce results with great accuracy as we found in this study. This suggests an important novel role for the peer community to share their lived experiences in helping in training and updating of semi-supervised models.

In terms of content, we found that personal tips and inspiration were the most widely shared categories in terms of volume and reach. This makes sense given the hashtag driving this content as *#mytipsformentalhealth* and reflects a wealth of peer support and personal insights from the mental health community. With nearly 45,000 Twitter posts in these two categories, there is a plethora of valuable information. Beyond raising awareness, the significant volume of this content raises the potential of matching people to the most relevant and useful messages. Although our methods focused on classification of content, we propose a need for new methods to help deliver the most relevant peer support messages to each individual as a next step in increasing the utility and potential of social media for mental health. Matching the right content to the right person could increase the impact of social media for health promotion [23], although will rely on accurate classification of content as outlined in this article.

Of concern, we found that stigmatizing content was most engaging in terms of retweets and favorites per Twitter post. Although the volume of stigmatizing content was relatively small compared to other categories, the higher social diffusion of stigmatizing content represents an opportunity for the mental health community to combat it. Either in the form of public health educational campaigns, direct outreach and education to those posting stigmatizing content, or working with social media platforms to remove harmful or hateful content—it may be possible to curb the spread of this content early and thus preclude its broader diffusion.

Further, that stigmatizing and inspirational posts using the awareness hashtag receive the highest amount of engagement from an audience on Twitter has implications to mental health awareness campaigns in two forms. First, concerning individuals on Twitter undergoing mental health challenges, reading about others' experiences (both stigmatizing and inspirational ones) might lead to a space for shared experiences, reducing stigma and pushing them to open up about their own experiences. To this end, it would be interesting to observe whether individuals with certain mental illnesses versus others are less or more positively influenced by exposure to such content. On the other hand, concerning individuals who might not have any mental health challenges, engaging with such content on their Twitter network, might lead to increased mental health literacy, awareness around experiences relating to mental health challenges and reduced stigma. More broadly, how discussion of stigma issues on social media affects mental health outcomes, both for those with and without lived experiences, constitutes an interesting direction for future research.

Next, we observed that across all categories, engagement received (per post) by favorites is higher than that received by retweets. This demonstrates that with regard to content related to mental health awareness campaigns, endorsement and acknowledgment behaviors as characterized by favorites are higher than information sharing and broadcasting behaviors characterized by retweets. This poses a challenge to online mental health awareness campaigns whose primary goal is to initiate information sharing and diffusion behaviors across a large network of individuals.

We also found an overall lack of resources and medical categorized posts compared to other categories. Although local resources will vary based on location, online resources such as the National Alliance of Mental Illness website are easily accessible to most people, if not all [49]. Promoting and introducing more resource-based content into social media discussions represents an opportunity for reaching new audiences and raising awareness of existing services. The paucity of medical-based Twitter posts is also interesting as it may represent a lack of engagement by the medical community in social media discussions. Although offering medical advice over social media is dangerous and unethical, information on the overall benefits of engaging in treatment, adhering to medications, and maintaining healthy weight and diet are universal medical messages that have a role in mental health discussions like those happening on Twitter. The current lack of medical content may reflect clinicians are concerned about engaging publicly on social media but may also represent a missed opportunity to partake in the modern dialogue around mental health. Together, our findings allow drawing a variety of qualitative inferences as well as frame hypotheses, which can be tested in future research.

Like all studies, ours study has several weaknesses that must be taken into consideration, however, many of which outline directions for future research. First, we only study a single hashtag and campaign, and it is unclear how our results may generalize to a different mental health campaign that may gain significant traction on social media. Similar considerations of (a lack of) generalization exist for other types of social media content and platforms as well—here we focused on Twitter, English language posts, and analyzed the textual content of posts; however, other campaigns may be more predominant on other platforms (e.g., Facebook) and additional modalities may be adopted for information dissemination and raising awareness (e.g., via images or videos). Location analysis of these social media posts, whether Twitter or another platform, was also beyond the scope of this work, but in future work can provide additional rich insights into how engagement and content sharing across different topical categories varied over geography. We also note that, although millions of individuals use social media, we cannot make the claim that insights gleaned from a framework like ours, while indicating public sentiment around mental health, may not be a true reflection of society [50, 51]. However, the focus of this work was to specifically examine what sentiments characterize social media campaigns about mental health, hence our findings should be interpreted with that specificity in mind. Further, we worked with the creator of this campaign as an author on the paper to ensure our interpretations reflected both the challenges and successes of this campaign. Second, although we had expert review and classification of Twitter content and involved directly with the campaign's creator, we acknowledge that there is no gold standard of consensus for classification of mental health content relating to campaigns. Future work can bolster our approach by additionally surveying or interviewing those who are active and prominent contributors in an online campaign. Third, our machine learning classifiers have certainly provided a mechanism to automatically scale analysis and understanding of social media content surrounding such campaigns; however, the performance metrics in the current classifier do imply additional room for improvement. We caution against conclusions that would directly use such a classifier to understand people's perceptions and attitudes around mental health, such as for infodemiology or infoveillance purposes [52], without any human or expert involvement and intervention. Fourth and finally, our work is essentially a case study of a "natural" mental health campaign, however, mental health promotion efforts may also motivate non-profits and other stakeholder agencies and individuals to conceive and launch targeted social media campaigns. To this point, it is important to appreciate that the social media environment is highly dynamic and rapidly evolving. Therefore, although our research does not provide a mechanism to assess the return on investment of these targeted campaigns, future work is needed to develop quantifiable and objective measures of social media campaigns, that takes into account their unique attributes, which are often absent in campaigns that are launched and driven offline. Important next steps for the field will be investigations comparing and contrasting different types of mental health campaigns released both on social media and others via other mechanisms.

CONCLUSION

Social media holds both promise and pitfalls for mental health. Although our results highlight the sheer volume of public discourse happening online today, challenges in accurately classifying this mental health content present barriers to fully using this information to inform local public health campaigns or allocate resources. Although new methods will help overcome these barriers, partnering with the peer community who is already creating and partaking in these online campaigns offers the next step for the field to better study and support the new public discourse on mental health.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Translational Behavioral Medicine* online.

Acknowledgments: K.S., S.K.E., and M.D.C. were partly supported by a National Institutes of Health grant R01GM112697. J.T. is supported by a research fellowship from the American Psychiatric Association Foundation.

Compliance with Ethical Standards

Conflict of Interest: All authors declare that they have no conflicts of interest.

Human Rights: For this type of study, formal consent is not required.

Informed Consent: This study does not involve human participants and informed consent was therefore not required.

Welfare of Animals: This article does not contain any studies with animals performed by any of the authors.

References

1. World Health Organization. 2018. Mental disorders. [online] Available at http://www.who.int/en/news-room/fact-sheets/detail/mental-disorders Accessed 10 June 2018.
2. Vigo D, Thornicroft G, Atun R. Estimating the true global burden of mental illness. *Lancet Psychiatry.* 2016;3(2):171–178.
3. Kilbourne AM, Beck K, Spaeth-Rublee B, et al. Measuring and improving the quality of mental health care: A global perspective. *World Psychiatry.* 2018;17(1):30–38.
4. Kietzmann JH, Hermkens K, McCarthy IP, Silvestre BS. Social media? Get serious! Understanding the functional building blocks of social media. *Bus Horiz.* 2011;54(3):241–251.
5. Kaplan AM, Haenlein M. Users of the world, unite! The challenges and opportunities of Social Media. *Bus Horiz.* 2010;53(1), 59–68.
6. Grajales FJ III, Sheps S, Ho K, Novak-Lauscher H, Eysenbach G. Social media: A review and tutorial of applications in medicine and health care. *J Med Internet Res.* 2014;16(2):e13.
7. Statista. 2018. Number of social media users worldwide 2010–2021 | Statista. [online] Available at https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/ Accessed 11 June 2018.
8. Birnbaum ML, Ernala SK, Rizvi AF, De Choudhury M, Kane JM. A collaborative approach to identifying social media markers of schizophrenia by employing machine learning and clinical appraisals. *J Med Internet Res.* 2017;19(8):e289.
9. Saha K, Weber I, Birnbaum ML, De Choudhury M. Characterizing awareness of schizophrenia among facebook users by leveraging facebook advertisement estimates. *J Med Internet Res.* 2017;19(5):e156.

10. Haskins BL, Davis-Martin R, Abar B, Baumann BM, Harralson T, Boudreaux ED. Health evaluation and referral assistant: A randomized controlled trial of a web-based screening, brief intervention, and referral to treatment system to reduce risky alcohol use among emergency department patients. *J Med Internet Res.* 2017;19(5):e119.
11. Luxton DD, June JD, Fairall JM. Social media and suicide: A public health perspective. *Am J Public Health.* 2012;102(suppl 2):S195–S200.
12. Moorhead SA, Hazlett DE, Harrison L, Carroll JK, Irwin A, Hoving C. A new dimension of health care: Systematic review of the uses, benefits, and limitations of social media for health communication. *J Med Internet Res.* 2013;15(4):e85.
13. Cassa CA, Chunara R, Mandl K, Brownstein JS. Twitter as a sentinel in emergency situations: Lessons from the Boston marathon explosions. *PLOS Curr Dis.* 2013. 1st edn. doi:10.1371/currents.dis.ad70cd1c8bc585e9470046cde334ee4b
14. Booth RG, Allen BN, Bray Jenkyn KM, Li L, Shariff SZ. Youth mental health services utilization rates after a large-scale social media campaign: Population-based interrupted time-series analysis. *JMIR Ment Health.* 2018;5(2):e27.
15. Freeman B, Potente S, Rock V, McIver J. Social media campaigns that make a difference: What can public health learn from the corporate sector and other social change marketers? *Public Health Res Pract.* 2015;25(2):e2521517.
16. Neiger BL, Thackeray R, Van Wagenen SA, et al. Use of social media in health promotion: Purposes, key performance indicators, and evaluation metrics. *Health Promot Pract.* 2012;13(2):159–164.
17. Kuehn BM. Twitter streams fuel big data approaches to health forecasting. *JAMA.* 2015;314(19):2010–2012.
18. Jashinsky J, Burton SH, Hanson CL, et al. Tracking suicide risk factors through twitter in the US. *Crisis* 2014;35(1):51–59.
19. Coppersmith G, Dredze M, Harman, C. Quantifying mental health signals in Twitter. In: Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality (pp. 51–60). 2014.
20. McClellan C, Ali MM, Mutter R, Kroutil L, Landwehr J. Using social media to monitor mental health discussions—evidence from Twitter. *J Am Med Inform Assoc.* 2017;24(3):496–502.
21. Mowery D, Bryan C, Conway M. *Feature studies to inform the classification of depressive symptoms from Twitter data for population health.* 2017. *arXiv preprint arXiv*:1701.08229.
22. Berry N, Lobban F, Belousov M, Emsley R, Nenadic G, Bucci S. #WhyWeTweetMH: Understanding why people use twitter to discuss mental health problems. *J Med Internet Res.* 2017;19(4):e107.
23. Hawn C. Take two aspirin and tweet me in the morning: How Twitter, Facebook, and other social media are reshaping health care. *Health Aff (Millwood).* 2009;28(2):361–368.
24. Robinson P, Turk D, Jilka S. et al. Measuring attitudes towards mental health using social media: Investigating stigma and trivialisation. *Soc Psychiatry Psychiatr Epidemiol* 2019;54:51. doi:10.1007/s00127-018-1571-5
25. Joseph AJ, Tandon N, Yang LH, et al. #Schizophrenia: Use and misuse on Twitter. *Schizophr Res.* 2015;165(2–3):111–115.
26. Hswen Y, Naslund JA, Brownstein JS, et al. Online communication about depression and anxiety among twitter users with schizophrenia: Preliminary findings to inform a digital phenotype using social media. *Psychiatr Q.* 2018;89:569. doi:10.1007/s11126-017-9559-y
27. Gulliver A, Griffiths KM, Christensen H, Mackinnon A, Calear AL, Parsons A, Stanimi rovic, R. Internet-based interventions to promote mental health help-seeking in elite athletes: An exploratory randomized controlled trial. *J Med Internet Res.* 2012;14(3):e69.
28. Wright KB. Researching Internet-based populations: Advantages and disadvantages of online survey research, online questionnaire authoring software packages, and web survey services. *J Comput-Mediat Commun.* 2005;10(3):JCMC1034.
29. Shanahan M. 2018. #MyTipsForMentalHealth is trending on Twitter and people actually gave really valuable advice. [online] BuzzFeed. Available at https://www.buzzfeed.com/morganshanahan/people-are-sharing-their-tips-for-maintaining-mental-heath Accessed 20 December 2018.
30. Shanahan M. 2018. People are sharing their tips for maintaining mental health on Twitter and it's honestly so inspiring. [online] BuzzFeed. Available at https://www.buzzfeed.com/morganshanahan/people-are-sharing-tips-for-maintaining-mental-heath?utm_term=.riNkAajZnN#.puad6QxErY Accessed 11 June 2018.
31. Cha M, Haddadi H, Benevenuto F, Gummadi PK. Measuring user influence in twitter: The million follower fallacy. *ICWSM.* 2010;10(10–17):30.
32. World Health Organization. 2018. World Mental Health Day—10 October [online] BuzzFeed. Available at https://www.who.int/mental_health/world-mental-health-day/en/ Accessed 20 December 2018.
33. Brown BB. *Delphi process: A methodology Used for the Elicitation of Opinions of Experts (No. RAND-P-3925).* Santa Monica, CA: RAND Corp.; 1968.
34. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. *IJCAI.* 1995;4(2):1137–1145.
35. Saha K, Chan L, De Barbaro K, Abowd GD, De Choudhury M. Inferring mood instability on social media by leveraging ecological momentary assessments. *Proc ACM Interact, Mob, Wearable Ubiquitous Technol.* 2017;1(3):95.
36. Chapelle O, Scholkopf B, Zien A. 2009. Semi-supervised learning (chapelle, o. *et al.*, eds.; 2006)[book reviews]. *IEEE Trans Neural Netw..* 2017;20(3):542–542.
37. Liu, J, Zhao S. Wang G. SSEL-ADE: A semi-supervised ensemble learning framework for extracting adverse drug events from social media. *Art intel med.* 2018;84:34–49.
38. Zhou ZH. When semi-supervised learning meets ensemble learning. In: *International Workshop on Multiple Classifier Systems* (pp. 529–538). Springer, Berlin, Heidelberg. 2009.
39. Muja M, Lowe DG. Fast approximate nearest neighbors with automatic algorithm configuration. *VISAPP (1).* 2009;2(331–340):2.
40. Eisenstein J, Ahmed A, Xing EP. Sparse additive generative models of text. In: *Proceedings of the 28th International Conference on International Conference on Machine Learning.* Omnipress. 2011:1041–1048.
41. Golder SA, Macy MW. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science.* 2011;333(6051):1878–1881.
42. Enli G. Twitter as arena for the authentic outsider: Exploring the social media campaigns of Trump and Clinton in the 2016 US presidential election. *Eur J Commun.* 2017;32(1):50–61.
43. Hanna R, Rohm A, Crittenden VL. We're all connected: The power of the social media ecosystem. *Bus Horiz.* 2011;54(3):265–273.
44. Shirky C. The political power of social media: Technology, the public sphere, and political change. *Foreign Aff.* 2011;90(1):28–41.
45. Guo C, Saxton GD. Tweeting social change: How social media are changing nonprofit advocacy. *Nonprofit Volunt Sect Q.* 2014;43(1):57–79.
46. Korda H, Itani Z. Harnessing social media for health promotion and behavior change. *Health Promot Pract.* 2013;14(1):15–23.
47. Insel T, Cuthbert B, Garvey M, Heinssen R, Pine DS. Research domain criteria (RDoC): Toward a new classification framework for research on mental disorders. *Am J Psyc.* 2010;167:7.
48. Taylor SE, Brown JD. Illusion and well-being: a social psychological perspective on mental health. *Psychol Bull.* 1988;103(2):193.
49. Hollis C, Morriss R, Martin J, et al. Technological innovations in mental healthcare: Harnessing the digital revolution. *Br J Psychiatry.* 2015;206(4):263–265.
50. Chou WY, Hunt YM, Beckjord EB, Moser RP, Hesse BW. Social media use in the United States: Implications for health communication. *J Med Internet Res.* 2009;11(4):e48.
51. Tufekci Z. Big questions for social media big data: Representativeness, validity and other methodological pitfalls. *ICWSM.* 2014;14:505–514.
52. Eysenbach G. Infodemiology and infoveillance: Framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the Internet. *J Med Internet Res.* 2009;11(1):e11.
53. Zhu X. Semi-supervised learning literature survey. *Comput Sci (University of Wisconsin-Madison).* 2006;2(3):4.
54. Friedman JH, Bentley JL, Finkel RA. An algorithm for finding best matches in logarithmic expected time. *ACM Trans Math Softw.* 1977;3(3):209–226.
55. Bentley JL. Multidimensional binary search trees used for associative searching. *Commun ACM.* 1975;18(9):509–517.
56. Paul MJ, Dredze M. Drug extraction from the web: Summarizing drug experiences with multi-dimensional topic models. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 168–178), 2013.
57. Sharma E, Saha K, Ernala SK, Ghoshal S, De Choudhury M. Analyzing ideological discourse on social media: A case study of the abortion debate. In: Proceedings of the 2017 International Conference of The Computational Social Science Society of the Americas(p. 3). ACM, 2017, October.
58. Chandrasekharan E, Pavalanathan U, Srinivasan A, Glynn A, Eisenstein J, Gilbert E. You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. *Proc ACM Hum-Comput Interact. 1(CSCW).* 2017;31.
59. Laniado D, Mika P. Making sense of twitter. In: International Semantic Web Conference (pp. 470–485). November, 2010; Berlin, Heidelberg: Springer.
60. Zhou Y, Zhan J, Luo J. Predicting multiple risky behaviors via multimedia content. In: International Conference on Social Informatics (pp. 65–73). September, 2017; Cham: Springer.
61. De Choudhury M, Kiciman E, Dredze M, Coppersmith G, Kumar M. 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 (pp. 2098–2110). ACM, 2016, May.
62. Schwartz HA, Eichstaedt JC, Kern ML, et al. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLoS One.* 2013;8(9):e73791.

63. Saha K, De Choudhury M. Modeling stress with social media around incidents of gun violence on college campuses. *Proc ACM Hum-Comput Interact 1(CSCW)*. 2017;92:1–92:27.
64. Pennebaker JW, Francis ME, Booth RJ. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*. 2001;71(2001):2001.
65. Nagatsuka T, Miyachi T, Shimada A, Takeya K, Kemmochi E, Nakajima A, Fujita, K. *U.S. Patent No. 7,194,471*. Washington, DC: U.S. Patent and Trademark Office.2007.
66. Calado P, Cristo M, Moura E, Ziviani N, Ribeiro-Neto B, Gonçalves MA. Combining link-based and content-based methods for web document classification. In: Proceedings of the Twelfth International Conference on Information and Knowledge Management (pp. 394–401). ACM, 2003, November.
67. Manning C, Surdeanu M, Bauer J, Finkel J, Bethard S, McClosky D. The Stanford CoreNLP natural language processing toolkit. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations (pp. 55–60), 2014.
68. Chung C, Pennebaker JW. The psychological functions of function words. *Soc Commun*. 2007;1:343–359.
69. Marcus MP, Marcinkiewicz MA, Santorini B. Building a large annotated corpus of English: the Penn Treebank. *Comput Linguist*. 1993;19(2):313–330.
70. Loader BD, Vromen A, Xenos MA. The networked young citizen: social media, political participation and civic engagement. *Inf Commun Soc*. 2014;17(2):143–150.