# #Bgnow on @Twitter: A Characterization Study of #Diabetes through #BigData


## ABSTRACT

Social media is continually used as a platform for sharing, seeking, and discussing around health challenges, transforming these platforms as a source for knowledge, support, and engagement for patient living with chronic diseases. Studying behavioural characteristic and the language of users online can offer invaluable insights on how health consumers interacts and influence each other on social networks. In this paper, we investigate to characterize the differences between language and behaviour of diabetic users on Twitter microblogging platform. In particular, we aim to investigate the behavioural distinction between patients who can successfully manage their diabetes and those who fail. We study patient behavioural online in terms of linguistic, textual and visual attributes and contents in their online posts. We have observed several characteristics such as negative affective, seeking and sharing supportive contents, and difference in shared visual concepts, which differs adopted and non-adopted users. We discuss the implication of our finding in providing better healthcare intervention and provide a supervised model which can predict the success of user based on his published online content.



## Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval-clustering

## General Terms

Algorithms, Performance, Experimentation

## Keywords

Health informatics


## 1. INTRODUCTION

Social media has been considered as a prominent means for seeking, sharing and spreading information, where users discuss about a verity of events and topic of their interests. In such a context, health consumers increasingly utilize social platforms to fulfill their health demands through seeking and sharing health information and experiences as well as providing online social support for their peers [4, 13, 14]. For example, it was reported that 57% of patients with chronic conditions constantly and actively refer to social media to acquire health information while 20% of them have already participate in generation of online health contents [16]. In essence, social environments and virtual communities have been transformed to a confident environment permitting users to be connected with their peers who have experienced similar conditions, difficulties, and challenges, assisting them to cope with their situations [9, 20]. The emerging of self-tracking gadgets and the enthusiasm of users in taking informed health decisions have also intensified this trend. This motivates users to not only share about their daily life events but also disclose their health information in social platforms [12]. For example diabetic patients frequently posts about their health conditions, medications, and the outcome of medications on social media platforms like Twitter and Instagram. Further, the ubiquity of social media encourages health consumers to not only discuss about their health conditions and share experiences but more importantly share their health related measurements, like blood pressure and blood glucose, which provides an invaluable resource to study and analysis individual's and communities' wellness and behaviour. While Electronic Health Records (EHRs) are increasingly utilized in medical informatics as an important and distinct data source, limited research efforts have been devoted into utilizing PGWD available on social networks [8, 39, 43].

Diabetes is the sixth leading cause of death in the US and it is estimated to be the seventh cause of the worldwide by 2030 [1]. It causes serious complications and can lead to poor quality of life [3]. People with diabetes are more suspectable to other illnesses. However, they can reduce the occurrence of these potential complications through diabetes self-management. Training and education of self-management of diabetes prevent unnecessary health care utilization and hospitalization and improve patients wellness [21]. However, diabetes self-management is not a trivial task for chronic patients since it is highly linked to several individual and cultural factors such as demographic attributes, psychological mood, patient knowledge and

[1] http://www.cdc.gov/diabetes/data/statistics/2014statisticsreport.html

lifestyle. The prominent role of self-management in diabetes has resulted into the emergence of several diabetes management programs and schemes [21]. However, the success and efficacy of this programs are still unclear or partially opaque due to the following reasons [18, 30]. First, most of these studies have been performed in a controlled clinical setting and only investigate some of health factors of interest like diet and activities. Even though these factors are important the effects can vary from patient to patient making the reported results almost restricted to the local community studied. Second, the adherence of the user is a major factor in these program. However, users adherence and ambitious vary from time to time making the generalization of result optimistic. Third, traditional studies in health sciences are mostly based on a observational study or based on survey data from patients which are limited in the number of subjects and the time period of the study, making it hard to derive a comprehensive conclusion. Therefore, the factors affect on success of users in managing their diabetes are still not fully studied. In particular, the role of personal characteristics like user's behavior and mood is still unclear.

In this paper, we investigate the success of users to self-manage their diabetes, the role of users' behavior on self-management, and how exactly this behavior correlates to self-management and health of user.In particular, we investigate the factors which differentiate two cohorts of users: Adopted and Non-Adopted diabetes patients, where they are divided by their success in controlling their blood glucose values on healthy target range. In contrast to traditional studies, we utilize users self-mention values of blood glucose on Twitter microbloging service. We study users from three aspects, namely linguistic attributes, published textual content, and shared multimedia contents, i.e., images on Instagram, to find their behavioural distiction. Specifically, we address the following research questions:

**RQ1:** What are the characteristics of these communities in terms of linguistic, behavioural, cognitive and affective styles and attributes?

**RQ2:** How do they differ in terms of their posts content published in social platforms?

**RQ3:** How different are their multimedia interests on social platforms, i.e. what entities and concepts do they often share in pictures on social networks?

Studying behavioural characteristics of successful patients provides significant insight about individual's wellness. First, understanding behaviour factors correlated with success in managing diabetes will help us to design proper intervention program to assist users coping with their condition and improve their wellness. Second, by understanding success factors in self-management, we can assist new diagnosed users as well as users struggling with their condition to adopt their lifestyle with the disease condition. Third, several years of research in user behaviour and wellness have clearly revealed that social networks and health are closely related through the social reinforcements come from observing and modeling others' behaviors [40]. By linking users in different categories, i.e. suggesting successful users to strugglers, we can assist them to learn positive behaviours from new friend on social communities [33, 27, 22]. This study is also beneficial for understanding collaborative behaviour of communities. By aggregating behaviour of individuals, we can shed light on public health, trying to understand the wellness patterns and trends of user groups in large scale. Overall, the answers to these questions will enhance our understanding of users' online interactions about diabetes, in particular on the Twitter microblogging platform and Instagram multimedia sharing service, and how their behaviours are correlated with their health condition, assisting us in various applications like group wellness analysis [44], public health [22, 31], and health policy [35], to name a few.

The contribution of this paper are threefold:

- We demonstrate an effective approach for collecting patient generated wellness data from social media, which scale well. In particular, we introduce an approach for harvesting wellness information available on social media for studying diabetes and especially diabetes success.
- We conducted a behavioural analysis of social media users who can successfully manage their diabetes and the factors which cause a fail.
- We conduct an open vocabulary analysis that captures language use of diabetic patients toward identifying behavioural patterns result in self-managing diabetes.
- We investigate media contents shared by diabetes patients to reveal their preferences in sharing various visual concepts.

Our results open a new research direction to study wellness in large scale utilizing social media data. Further, the study demonstrates the potential of social media to design proper intervention and treatment programs for diabetic patients. The remainder of this paper is organized as follows. In Section 2, we review related studies which utilize social media for wellness. We next explain data collection approach and grand-truth construction in Section 3. Section 4 describes several analytical experiments we conducted in this paper followed by experimental results in Section 5. In Section 6, we conduct a success prediction study for diabetic patients followed by discussion in Section 7. Finally, we summarize the paper and outline the future work in Section 8.

# 2. RELATED WORK

# 3. DATA COLLECTION AND GROUND-TRUTH GENERATION

## 3.1 Dataset

We harvested public data available on social media for the construction of dataset as well as ground truth labels. We utilized a set of diabetic users who actively share their wellness information on Twitter and Instagram. These users, besides posting about general topics and events, actively post their lifestyle information and activities

**Table 1: Statistics of the Dataset**

| | Twitter | Instagram |
|---|---|---|
| # of Users | 649 | 113 |
| # of Posts | 1,507,681 | 2,623 |
| # of Bgnow Posts | 20,079 | – |
| Max Degree of the Users | 7,695 | - |
| Min Degree of the Users | 1 | - |
| Avg. Posts per Users | 2,323 | 17 |

such as their diet, activities, and emotional states. They further disclose their health information in terms of medical events and measurements like the onset of hypoglycaemia/heyperglocemia and their blood glucose values, respectively. A main characteristic of this cohort of users is that they report the exact values of their blood glucose in tweets using '#BgNow' hashtag. This hashtag is popular among diabetic patients to report their blood glucose values and other diabetic related information online. In essence, '#BgNow' plays the role of an online support group for diabetic patients through which they share, explore, and validate their health states and knowledge. The hashtag is special interest for us since it enables us to study the correlation between diabetes patients behaviours and their health indicators, e.g. blood glucose value. Indeed, #bgnow hashtag acts as a social sensor through which we can measure the blood glucose values of individuals on social networks.

Analyzing this collection of tweets provides invaluable insight about diabetic patients and their health information disclose on social media as well as the correlation between their behaviour online and their health conditions. We next collected all the published tweets in the period of six months from May to November 2015 resulting into more than three million tweets. We next filter out all the users who did not post any status for one month or post few tweets. To link the twitter account of users with their Instagram, we obtained the cross-platform links in which they publish a post from their Instagram account on Twitter. By storing all these link, we can find their Instagram account and avoid the problem of user identification across different networks[47]. It is worth noting that, by utilizing this approach, we may fail to find the corresponding Instagram for those users who are not post cross network, resulting into a smaller datast for Instagram. Table 1 shows statistics of our datasets for both Twitter and Instagram.

Table 2 shows several representative tweets from the dataset. As you can see, there are three major tweet categories pertaining to diet, activity, and health information of individuals. It is worth noting that except these tweets, there are other tweets which may not directly related to health of individuals as demonstrated in the last row of Table 2.

## 3.2 Ground-truth Generation

Medical studies suggests that diabetic users maintain their blood glucose between 70 to 130, which is considered as controlled blood glucose. A measurement between this range demonstrates that the patient could successfully manege his blood glucose while he was unsuccessful otherwise. The out of range values normally need to be corrected by lifestyle

**Table 2: Examples of wellness tweets from dataset**

| |
|---|
| Diet |
| Too much drink in party<br>Talking about hot chocolates, I might just go and make myself one :D<br>found Taylor's pretzels in my backpack and I'm so happy wow<br>Eat 20g carbs and go fo running |
| Exercise |
| after 1 hour run #bgnow 130<br>I just finished 1 hour biking<br>BGnow 95, thanks swimming pool |
| Health |
| Feel too much Fatigue<br>ate great oatmeal, toast, and eggs. Had 1 unit insulin |
| Others |
| blood sugar taking a nosedive! was it the symlin or the 10 mile run?<br>Can diabetics eat plums?: URL URL<br>A low sugar diet, I eat fruit? Some bad or better for me? URL |

changes or treatments. Motivated by this principle, we partition the users into two distinct cohorts based on their reported blood glucose values: **Adapted Cohort** (**AC**), and **Non-adapted Cohort** (**NC**). The definition of these two cohorts are as follow,

- **Adopted Cohort (AC)**: An adapted user is able to control and maintain his values of blood glucose in the suggested range most of the times. For such a user, his blood glucose measurements are in the suggested range more than $t\%$ of times.
- **Non-adopted Cohort (NC)**: A user is non-adopted if he fail to have a controlled blood glucose, i.e., his blood glucose measurements are usually out of range.

We set the threshold $t = 50$ to divide the users in our dataset into two groups of different blood glucose patterns. Mathematically, we utilized the following decision function to construct ground truth labels,

$$d(u_i) = \begin{cases} +1 & \text{if } \frac{Pr(u_i \in AC)}{Pr(u_i \in NC)} \geq t \\ -1 & \text{otherwise} \end{cases}, \quad (1)$$

where $Pr(u_i \in AC)$, and $Pr(u_i \in NC)$ are the probability that the measurements for user $u_i$ are in the controlled range and out of the controlled range, respectively. This grouping is coarse, but it is motivated by health studies [17, 34] stating that users who can manage their blood glucose will have a better long term health and fewer diabetes complications. In the future, we aim to define more detailed groups, e.g., users who have on target, below target, above target measurements and also different trends like stable, and fluctuating trends. By using Eq.(1), we can intuitively divide our diabetic patients into two groups of users, which clearly show how they have managed their diabetes. In the remainder of the paper, we study how these two communities' online behaviour differ in terms of linguistic, textual, and visual content published on their social network posts.

## 3.3 Attribute Extraction

The aim of our research is to investigate the behavioural distinction between two different cohorts of patients with respect to their health attributes, in our study blood glucose values. To do so, we need to extract the reported

**Table 3: Representative examples of regular expressions for extracting blood glucose values from users' posts.**

| | | |
|---|---|---|
| 195.0 BG @ 08:20AM after bike ride 90 minutes | NUMBER (mg\|mmol) (BG\|Bgnow) (at\|@) TIME * | BG : 195 Time: 08:20 |
| Going on a 3 mile run, #BGnow 120 #bigbluetest | * (BG\|BGnow) NUMBER | BG : 120 |

measurements of blood glucose values in their tweets. Several approaches have been proposed in information extraction to detect the right piece of information from a text corpus, like pattern-based methods, supervised classification, and Conditional Random Field (CRF) [6]. Here, we utilized a simple but effective rule based approach. Intuitively, we defined a set of regular expression to extract the measurement values of blood glucose for individuals as shown in Table 3. We followed a bootstrapping approach of [46] to ensure the coverage and diversity of used patterns, where all extracted patterns are manually verified to ensure accuracy. Given a user post, we apply these set of rules to find whether a given tweet contains any reported valued of blood glucose[2].

# 4. CONTENT ANALYSIS

In this section, we applied different analytical experiments we applied for understanding the behavioural distinction between AC and NC users.

## 4.1 LIWC Analysis

To identify and understand behaviour distinction of adapted and non-adapted groups, we leverage a variety of indicators including linguistic and non-linguistic indicators. The motivation behinds the investigation is that several psychological studies demonstrate behavioural expression of individuals and their responses expose their life context, crises, and vulnerabilities [38]. Our analysis, in this section, is largely based on LIWC, which has been widely used in literature to study individuals behaviours in depression [11], addiction recovery [32], anorexia [10], to name a few. We hence examine three categories of attributes named (1) affective attributes, (2) cognitive attributes and (3) linguistic and stylistic attributes.

*Affective attributes.* Affective measures have attracted a lot of research in text and opinion mining to detect the objectivity of user towards products, organization, and services [29]. Recently, affective measures have been largely utilized to measure emotional disclosure of users in social media [28]. Motivated from prior literature, we measure positive affect (PA) and negative affect (NA) based on LIWC categories. We also compute four other emotional expression indicators: anger, anxiety, sadness, and swear.

*Cognitive attributes.* Several studies in psychology have demonstrated that cognitive process is largely associated with health improvement. For example, greater usage of cognitive words is related to less anxiety after a treatment [1]. Cognitive words are also utilized for explanatory purposes and demonstrate the demands of individuals for understanding the situation. We therefore evaluate the cognitive process of individuals based on cognition and perception word categories of LIWC.

*Linguistic attributes.* We consider five measures of linguistic style: (a) Lexical Density: consisting of words that are verbs, adjectives (identified using NLTK's POS tagger), and adverbs. (b) Temporal References: consisting of past, present, and future tenses. (c) Social/Personal Concerns: words belonging to family, friends, social, work, health, and death. (d) Interpersonal Awareness and Focus: words that are 1st person singular, 1st person plural, 2nd person, and 3rd person pronouns. (e) We also evaluated words associated with quantities such as numeric values as diabetes patients frequently needs to consider amount of and quantities of their foods, medications and activities to manage their health condition.

## 4.2 Topicical Content Analysis

We studied the textual contents shared by diabetic users from two aspects: words and phrases (N-grams), and topics discussed.

*N-Gram Analytic.* In addition to linguistic analysis, we also investigate the usage of various n-grams in the contents shared by adopted and non-adopted people. Specifically, we investigate to discover the difference in usage of uni-, bi-, tri-grams between two groups[3]. However, comparison between two set of n-grams is a challenging task mostly demonstrated by word-cloud. Inspired by recent research works in computational social sciences [28], we compute the log-likelihood of the ratio between usage pattern of each n-gram between adopted and non-adopted groups. Mathematically, it can be computed as follows,

$$LLR = 2 \times [\ln\big(Pr(u_i \in AC)\big) - \ln\big(Pr(u_i \in NC)\big)] \quad (2)$$

Indeed, $LLR$ demonstrates a clear measure to compare the differences between usage of a n-gram between two groups. As it computes the log likelihood ratio, when a n-gram is equally used in two groups then its $LLR$ will be near zero. Meanwhile, it would be greater than zero if it is more frequent in first group in compare to the second group, whereas it would be less than zero if the pattern is reverse [4].

*Topic Analytic.* Although content analysis based on n-grams provides an intuitive way for understanding the published contents by individuals, it processes the text based on low level features, i.e., words, and fail to capture high level semantics inside the text. We hence apply topic models to discover the semantic topics inside the posts published by different user groups. Topic models have been commonly used to analyze health data [37]. Following the prior literature, we obtain topics by applying latent Dirichlet Allocation (LDA) over the entire set of posts shared by all users. To train the topic model, we used the default

[2]In some cases, several numerical values might be found as a candidate for blood glucose value. We used the value which is closer to #Bgnow hashtag as the reference value.

[3]In this paper, we use the general term 'n-gram' to refer to uni-, bi-, and tri-grams in the text.

[4]To compute LLR measure, we assume that all n-grams are probable in both cohorts with a very low prior probability of $p = 10^{-6}$.

hyper-parameter settings and set the number of topics to 50, which we observed to work well in our experiments [5]. To measure topic differences between two groups of users, we first compute the posterior probability of each topic separately for the adopted and non-adopted users. We then compute the rate of increase for each topic as the difference between the posterior of the topic using the $LLR$ measure, which is the difference between logarithm of the ratio of posterior probability of the topic in adopted group to non-adopted group.

## 4.3 Visual Content Analysis

Social platforms are heterogenous information networks including multi-modal contents. Visual contents has become essential part of today social interactions. Hence, we also investigate the differences between AC and NC groups according to the visual contents are shared in Instagram social service. Comparing visual concepts of shared images however is a challenging tasks due to the richness and complexity of shared images. To effectively represent visual contents, we represent each image with a bag-of-visual-concept in which each images is represented with a vector of visual concepts happening in it. Inspired by prior studies [], we utilized 1000 visual concepts of ImageNet as a predefined visual concept dictionary due to its popularity in multimedia studies []. We hence constructed a feature vector for each image based on the state-of-the-art deep learning architecture of GoogleNet []. We next compute the user feature vector by averaging the feature vector of all images which were shared by him.

# 5. RESULTS

## 5.1 LIWC Analysis

Table 4 summarizes the LIWC measures of behavioural attributes for the two cohorts of users, adopted and non-adopted users. Overall, the contents published by AC users are less negative than those published NC users, demonstrating that AC users have a positive prespective towards their health and lifestyle.

*Affective attributes.* As can be seen from the table, adopted users are less negative. Pervious studies also reported similar correlation where negative affection is associated with poor health conditions and engagement [41]. Further, NC contents demonstrate more anger and sadness rather than AC contents. This result may attribute to the fact that being unsuccessful to cope with their issues make patient to be more angry and feel hopeless, loneliness and restless. The impact can be amplified in reverse direction where feeling hopeless and loneliness is highly correlated with less engagement and success.

*Cognitive attributes.* In terms of cognitive attributes, AC patients use more negation structures such as 'not', 'no' as compared to NC patients. Further, they also share perception words, i.e. 'see' and 'feel', which shows they are more likely to express their feeling. Meanwhile, NC users use less certainty in their publishing which is associated to more self-consciousness rather than users who are able to control their health condition [45]. This finding is interesting where indicates NC users feel guilty regarding their situation and hence may engage less with their community (See section 5.2 for more results).

[5]We tuned the number if topics by perplexity as suggested by [48].

**Table 4: The result of $t$-test between posts published by AC and NC.**

| Category | AC | NC | p-value |
|---|---|---|---|
| **Affective** | | | |
| Positive | 4.278009 | 4.167514 | 0.018 |
| Negative | 1.580045 | 1.659568 | 0.042 |
| Anxiety | 0.268597 | 0.270568 | 0.032 |
| Anger | 0.268597 | 0.485459 | 0.012 |
| Sad | 0.268597 | 0.411838 | 0.094 |
| Swear | 0.137783 | 0.167541 | 0.345 |
| **Cognitive** | | | |
| Negation | 1.289593 | 1.315486 | 0.240 |
| Certainty | 1.032172 | 1.031162 | 0.121 |
| Cognition | 0.757557 | 0.750351 | 0.072 |
| Perception | 0.478281 | 0.465892 | 0.341 |
| **Linguistic style**: Lexical density | | | |
| Word Counts | 31638 | 38749 | 0.012 |
| Word Per Sentence | 21 | 156 | 0.034 |
| Verbs | 10.565475 | 10.669378 | 0.079 |
| Adjectives | 3.581538 | 3.635270 | 0.162 |
| Adverbs | 0.000021 | 0.000010 | 0.241 |
| **Linguistic style**: Tense | | | |
| Past tense | 1.925339 | 1.938270 | 0.374 |
| Present tense | 7.179502 | 7.307703 | 0.332 |
| Future tense | 0.736063 | 0.686216 | 0.009 |
| **Linguistic style**: Interpersonal awareness | | | |
| 1st person singular | 3.767330 | 3.967865 | 0.074 |
| 1st person plural | 0.559864 | 0.561946 | 0.046 |
| 2nd person | 1.427149 | 1.534811 | 0.064 |
| 3th person | 0.652806 | 0.72846 | 0.032 |
| **Linguistic style**: Quantities | | | |
| quantities+numbers | 6.072534 | 5.480919 | 0.008 |
| **Linguistic style**: Social concerns | | | |
| social | 5.903575 | 6.194108 | 0.034 |
| family | 0.250452 | 0.274595 | 0.073 |
| friend | 0.142805 | 0.159027 | 0.147 |
| health | 1.900407 | 1.957568 | 0.322 |
| death | 0.104887 | 0.131595 | 0.074 |
| work | 1.833484 | 1.678000 | 0.265 |

*Linguistic style attributes.* NC users have higher lexical density. They also share longer sentences as compared to AC. This is to be expected as NC users utilize social media as a means for acquiring information about their health concerns, as pointed by [13, 19]; such contents are mostly about self and hence people try to describe their situation completely [5]. This result needs to be studied more carefully as some studies associate lower lexical density to negative emotions as NC already has shown such characteristics. NC contents are more concerning about past and less focused about future, while AC users are more discussing about future. This is likely attributed to the anxiety of users and their concerns about their health conditions and issues. The literature has leveraged that lower future concerns is a known attribute of negative attitude towards user's own life, arising from their problems in managing their health condition[7]. Further, NC users show less social concerns since negative thought are associated with the self. Hence, they are less likely to talk about social concerns and community topics. More surprisingly, AC users are less concern about health and death as compared with NC users. This can be owning to the fact that these users have already adopted their

lifestyle to their situation and health condition; concerning less about their health and their disease consequences. Further, AC users have already shown positive affection in their behavior so they may less discuss about negative concepts like death. Last but not least, AC users are more talk about quantities and numbers; this is an important finding and specific to our study. It obviously shows that those are successful to manage their condition are concern about the quantities, which is deeply related to self-management of their condition. This demonstrates that diabetes management demands a careful consideration to balance the lifestyle; adjust their consuming calories, specifically carbohydrates, which will result into a successful management. This finding was already reported by research efforts in health sciences. However, our results reveal that people who discuss about quantities in social media are more likely to follow the correct management program. Overall it attests the potential value of social media as an intervention source for sciences as well as an information source for health studies.

## 5.2 Topical Content Analysis

In this section, we investigate how the contents published by these two cohorts of users are different from each other. As mentioned in section 4.2, we studied the frequency of different n-grams and topics in posts published by users.

*N-grams comparison.* We observed a great distinction between the usage of n-grams in posting of AC and NC users. Generally speaking, the content published by two groups of diabetic users demonstrates that AC users mainly acts as a social supporter or content providers for diabetic users. This finding itself is interesting and demonstrates that, by designing appropriate intervention tools/programs, we can help diabetic users to better manage their health condition.

Table 5 shows a list of 60 different n-grams organized in three groups with their associated LLR values. The top group lists n-grams with highest LLR values, demonstrating those are important for AC users; the second group shows those which are equally used by two groups of users; and the last group summarizes the list of important n-grams for NC users. Overall, our finding verifies the results observed in the last section, as we observe positive n-grams in the first group compared to the last group, which may attribute to the fact that NCs are struggling with their diabetes and focusing to find a proper way to manage their condition. In contrast, AC users are optimistic to the situation and spread positive emotions and experiences. The following contextual themes were observed from the data. (1) We found clear evidences of anxiety and anger (e.g., 'crazy sugar', 'shit', and 'hate') in NC contents owing to the fact that managing diabetes is problematic, in some sense, for this cohorts. This shows that online environment may be a place for them to release their emotional pressure through interacting with their peers. (2) Contents on seeking help and assistance ('how to', 'really want', 'suggestion') is also evident in NC users. This finding align with the pervious one which shows social platform may be perceived as a supporting environment for patients with diabetes, where not only users seek emotional support but also ask for informational support [49]. Several recent studies in computational social sciences also have attested that emotional support is a major motivation for chronic disease patients. Retrospective studies have reported that receiving emotional support is one of the main intentions that attracts users to utilize social networks for health, especially for chronic diseases like diabetes, insomnia, depression and so on [20, 23, 45]. (3) AC users, however, more frequently use positive words ('felt good', 'beautiful', 'nice', 'lovely'). The use of positive n-grams shows a positive view on the life and the tendency of spreading positive emotions and feelings. (4) Compared to NC users, AC users use diabetes management tool and platform like 'mysugar', demonstrating they are more curious and ambitious on managing their diabetes. Despite the importance and value of using computational framework in managing health problems, the benefits and impacts of using computational frameworks, from simple recording to high-end supporting framework like 'mysugar'[6], and 'onedrop'[7], in managing diabetes is still not fully investigated and more research needs to be investigated.

**Table 5: The result of n-gram study between posts published by AC and NC.**

| N-gram | $LLR$ | N-gram | $LLR$ |
|---|---|---|---|
| **N-grams (AC > NC)** | | | |
| miles hour | 7.222 | mins felt good | 7.179 |
| ran miles | 6.348 | strides | 6.348 |
| felt good | 1.953 | hills | 1.873 |
| keeping | 1.873 | ride | 1.801 |
| check strava | 1.448 | min walk | 1.284 |
| awesome | 1.251 | finish | 0.887 |
| sweatbetes | 0.873 | beautiful | 0.738 |
| a sweet life | 0.782 | cure | 0.732 |
| ready | 0.715 | lovely | 0.677 |
| mysugar | 0.642 | easy | 0.630 |
| **N-grams (AC ≃ NC)** | | | |
| blood sugar | 0.001 | diabetic | 0.001 |
| #dsma | 0.004 | pancreas | 0.004 |
| eating | 0.004 | insulin | 0.009 |
| injection | 0.009 | sugar | 0.009 |
| test blood sugar | 0.011 | take | 0.011 |
| regular | 0.011 | treatment | 0.013 |
| check | 0.013 | needles | 0.013 |
| fact | 0.013 | drink | 0.013 |
| insulin injection | 0.018 | health | 0.018 |
| control | 0.022 | injection site | 0.022 |
| **N-grams (AC < NC)** | | | |
| feeling support | −0.419 | num hr later | −0.419 |
| units novorapid | −0.419 | novorapid 2hr later | −0.418 |
| sugar level | −0.418 | continued ride cyclemeter | −0.418 |
| suggestion | −0.418 | glucose level | −0.418 |
| hate | −0.418 | shouldnt hurt bgnow | −0.418 |
| hurt bgnow | −0.418 | high | −0.418 |
| really want | −0.415 | latest level | −0.389 |
| weird | −0.388 | crazy sugar | −0.388 |
| stupid | −0.364 | shit | −0.361 |
| how to | −0.331 | nightmare | −0.301 |

*Topic analysis.* We also extract the underlying topics exists in the corpus and investigate to which extend the content published by AC and NC users are different from each others. We adopted LDA framework as described in Section 4.2 to extract textual topics from document corpus. Figure 1 depicts the differences of the existence of topics across two groups. As you can see from the Figure, the mean change across two groups is 20%, which

[6] `https://mysugr.com/`
[7] `http://onedrop.today/`

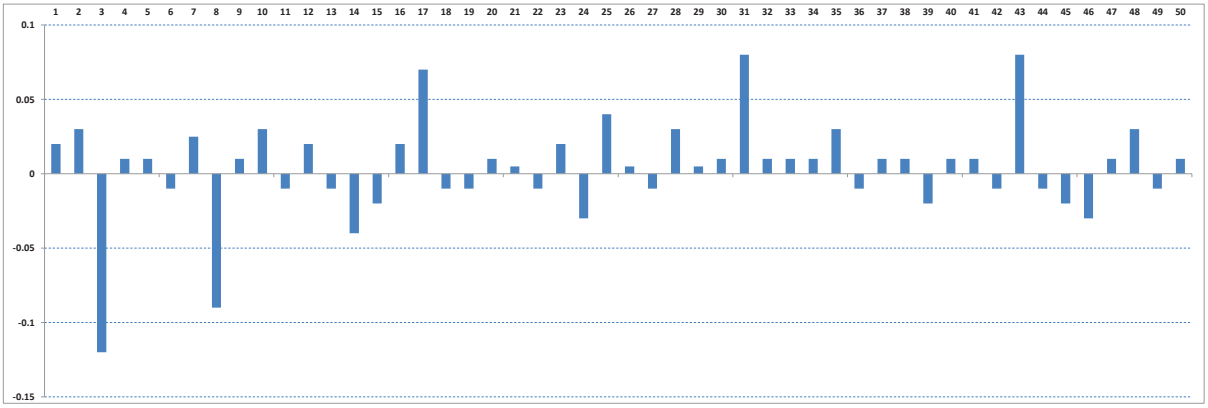
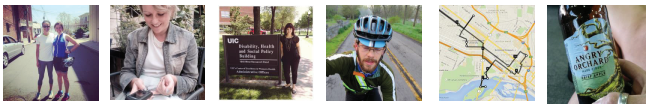



**Figure 1: Topics discussed by two cohorts of users**

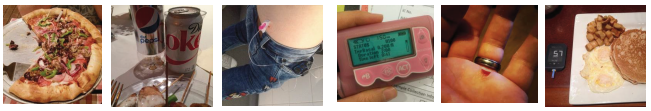
shows two cohorts of users are discussing on different topics online. Specifically, we observe that topics #3 (Self-critical), #8 (Conflicting feelings), #17 (Activities and Sports), #31 (Social support), #43 (Authorized information) show notable variations between two groups. To have a better sense of these topics we listed top 10 representative words from each topics in Table 6. From the Figure 1, following points can be observed. (1) Topic #3 represents 'self-critical' contents and the thoughts about being guilty and negative thought about self in NC cohort. This can be attributed into the failure of users in self-managing diabetes and similarly their desire to handle the situation. This is also consistent with the literatures, reporting that chronic disease sufferers develops tendencies of self-criticism which sometimes goes beyond the normal level and may result in mental disorders [2, 24]. (2) Topic #8 represents struggling and conflicting feelings and emotions that are often perceived by NC users. Investigating on the springs and outcomes of this conflicting emotions is worthwhile, which may result to establish better intervention in lifestyle medicine. (3) Discussing about activities and sports is common in AC communities as expressed by topic #43. Indeed a detail checking of extracted topics shows that two other topics which thematically talks about activities (#25, #28); however, they did not show strong distinction between two groups of diabetes. This result verifies the findings from health sciences which states the positive correlation between regular sport activities like 'running' and better management of diabetes, especially diabetes Type II[26]. It is worth noting that users discussing about sports and exercise activities in social media often utilize tracking devices linked with web portal and mobile applications, which assist them recording the history of their activities and planning for future. The finding is aligned with a recent research reporting that persistent usage of mobile applications significantly increase the success of users in weight loss program [36]. Despite the increasing popularity of tracker devices and mobile applications, limited studies have investigated their impacts and roles in managing chronic diseases, especially diabetes. (4) Topic #31 describes contents related to social support in online communities. Indeed, #Bgnow acts as a support groups, or a fast-response, support group for patient with diabetes, where they seek and provide informational and emotional support for their peers. This was already verified by the difference between n-grams usage within two

**Table 6: Examples of topics and corresponding representative words**

| ID | Representative Words | Topic |
|---|---|---|
| T3 | tired, hate, missed, hurt, horrible, struggled, sick, damn, ugh, lol | Self-critical |
| T8 | afraid, want, useless, comfortable, bad, pain, except, tough, nothing, easy | Conflicting feeling and emotions |
| T17 | running, gym, daily, hypo, mile, ride, walking, sugar, cyclometer, check | Activities and Sports |
| T31 | talk, dsma, advice, bed, insulin, diet, nutrition, sugar, ask, help | Social support |
| T43 | research, interested, diet, DiabeticDiary, prove, fact, fitness, food, sleep, hyperglycemia | Authorized information |



(a)



(b)

**Figure 2: Exemplar images have been shared by AC users (a), and NC users (b).**

groups. (5) Topic #17 reflects authorized information about diabetes, medications and management programs, showing that diabetic patients in AC groups spread more authorized information about diabetes. In essence, professional health providers leverage the power of social media to disseminate health content for health seekers and AC users republish this information in their network. This highlights another aspects of social media in healthcare.

## 5.3 Visual Content Analysis

Figure 2 shows some representative images from the images which have been shared by two groups of users. Our analysis demonstrates that the visual consents in shared images of users' Instagram accounts are highly related to the success of users in managing his diabetes. Figure 3 depicts top 20 statistically significant correlations between visual concepts

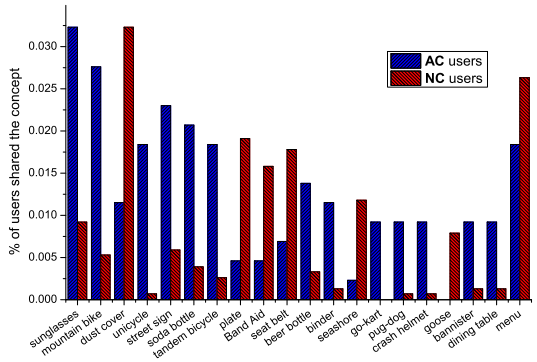



**Figure 3: Visual concepts which commonly shared by two cohorts of users: AC, and NC users.**

and the category of users. Several interesting signals can be observed from the Figure. For example, the visual concepts 'mountain bike' and 'unicycle' are positively correlated to AC category which demonstrates the strong preference of AC users to manage their diabetes with lifestyle change. Some other concepts like 'sunglasses', 'crash helmet', and 'street sign' which are objects related to activities are also demonstrate the same preferences. Retrospective research have also reported similar results in obesity and fitness related studies from social media platforms []. Conversely, visual concepts 'menu', and 'plate' are correlated to NC users, which illustrates the possible reason for failing to control their blood glucose level. Further, the positive correlation between the visual concept 'Band Aid' and NC users may indicate that they like to share the picture of their injection site, e.g. insulin injection or pump site, which shows their anxiety related to their health condition.

# 6. SUCCESS PREDICTION OF DIABETES MANAGEMENT

Success prediction is the task of predicting whether a specific user can successfully maintain his/her health indicators in a suggested range. In this paper, we aim to predict the success of a diabetic patient in controlling his blood glucose value, i.e., he/she belongs to the AC or NC group. In wellness domain, success prediction is an important task and many downstream decisions, such as interventions, treatments, and medications, are strongly related to the success of the patient in the current procedure/treatment. Here, we considered the success prediction as a binary classification problem in which we predict the category of users between AC and NC groups.

To accomplish this end, we use different behavioural attributes studied in last sections as individual features for training and evaluation in a supervised learning method. Each patient is represented with a vector of features and classified into one of two groups. We utilize a linear Support Vector Machine (SVM) as a binary classifier with default setting of scikit-learn [8]. For all experiments, we reported the average performance of 10 random experiments based on 10-fold cross validation. Table 7 demonstrates the performance

[8]http://scikit-learn.org/

**Table 7: The performance of different features settings for success prediction task**

| | Precision | Recall | F1 |
|---|---|---|---|
| LIWC | 59.57 | 62.46 | 60.18 |
| N-grams | 60.15 | 63.00 | 60.41 |
| Topics | 53.41 | 55.14 | 52.01 |
| All | 75.22 | 71.36 | 71.21 |

of different approaches in terms of precision, recall, and F1 metrics. To better understand the performance of each attribute, we reported the performance of each behavioral attributes individually and the aggregation of all features. As can be seen from the table, 'N-grams' outperformed the two other features 'LIWC' and 'Latent Topics'. This may be attributed to the fact that n-grams can better capture the semantic context of the text messages. 'Latent Topics' demonstrates the lowest performance which is expected since high level semantic topics fail to capture all aspects of text messages.

# 7. DISCUSSION

Our findings reveal several characteristics of social media, specifically Twitter, for diabetes. Many of our finding aligns with prior studies, stating that social media is a rich platform for health consumers through which they seek and share health information. Overall, they demonstrate that people online behaviour expose their health conditions and states as well as their success in adopting their life style to their wellness condition. Waving together these observation, it demonstrates that patient generated wellness data on social media can be effectively utilized for designing better intervention programs and services to assist patienta in better management of their diabetes.

## 7.1 Clinical Relevance

The abundant amount of available data can also assist us to better understand patient behaviours and detect potential issues resulting failures in self-management of diabetes. From a clinical perspective, social media can be utilized to complement patient self-report diaries by implicitly tracking his/her online behaviour. Social media can also assist to provide intervention through non-intrusive assessment of content providing and publishing, as discussed below.

**Persuasion Oriented Intervention**. With a proper lifestyle and behavioural change, we can successfully manage several chronic diseases such as diabetes, and obesity. While it seems an easy task, in practice, changing lifestyle is a challenging and complex task. According the Fogg Behavioural Model (FBM), three elements need to converge together in order to a behaviour occur: motivation, ability, and trigger [15]. Indeed, when a behaviour does not occur, one of these element is missing. Chronic disease sufferers usually have enough motivation to perform the target behaviour, which is suggested by the management program; however, they frequently will not trigger to perform the task on the correct time, e.g. reduce their sugar consumption or their sedentary lifestyle. By utilizing social networks not only we can understand the user lifestyle and wellness condition, but, more importantly, it is possible to motivate the user and trigger him in the same time. For example, we can suggest him some interesting outdoor activity based

on his past preferences or suggest him to have a more healthier meal. Further, providing useful health information regarding his/her health condition can effectively motivate user to follow the disease management program, in our case of the diabetes management program.

**Social Influence Intervention**. In psychology, social influence theory attests that individual's emotions, opinions, and behaviours are affected by others. Social influence has been studied in different domains and environments such as sales, marketing, leadership, and so on [25]. The holistic concept of wellness traditionally has been studied from different aspects includes, physical, mental, social, and spiritual components. Late studies extends these perspective to the social interactions finding that social interactions may affect individual's wellness either in positive and negative manner. For example, recent studies have revealed that person's circle of friends may influence his/her weight[42] and his/her sport activity level, i.e., how active he/she is in sports. Upon these findings, we can assist NC diabetics users to better manage their health condition through connecting them to AC users, i.e. diabetic users who already find how to successfully manage their disease.

## 7.2 Ethics and Limitations

While thinking about designing intervention programs on social network and in general health, it is important to bear in mind that wellness and health data can be extremely sensitive and need to be verified before providing to the user. Finding the authorized and reliable wellness information is a challenging task especially in a noisy platform such as social media with a lot of user generated contents and spams contents. The truthworthy of the information needs to be consider with a proper automatic or semi-automatic way. This can be done by a human in loop procedure to verify the potential risky unreliable information. Further, the design consideration in social platforms should honor the privacy of the affected individuals and abide the proper ethical guidelines ensuring that the intended profit of the intervention exceeds the potential difficulties and risks. To sum, we hope this research open a new avenue to not only detect and help diabetic patients through social platforms, but also understand the collaborative behaviour of different diabetics communities towards designing better healthcare interventions and treatments.

It is worthwhile noting that our paper does not make any claim to attributing the social network as an individual platform through which we can obtain a complete understanding of wellness condition of diabetic users and provide a full intervention program. We however attest that patient generated wellness and lifestyle data on social media can be utilized as a complementary source through which we can sense users wellness and lifestyle. Social sensors can be utilized as a complementary source of information in combination of the popular concept of quantified-self measuring users' attributes with wearable devices. We caution against using this method as standalone technique for diagnosis and prediction of diabetes. We also note that social media is a noisy and sparse platform where many users may not utilize it for health information explicitly; however, the implicit signals and clues in their social account can provide useful information when aggregated in scale.

Finally, our findings reveals the richness of patient generated wellness data on social media demonstrating that it can be used in combination of other information sources to obtain a comprehensive understanding of diabetes patient. It also raise several difficult questions for researchers, as mentioned below. How much social media information are reliable in health domain? How precise user's online behaviour reveal his offline attributes and behaviours? and How effective would the designed intervention be, in terms of changing user behaviour?

# 8. CONCLUSIONS AND FUTURE WORK

Social media is continually used as a platform for informational and emotional support around health challenges transforming these platform as a source for knowledge, support and engagement for patients living with chronic diseases such as diabetes. In such a context patients are encouraged to shared the exact values of their health measurements such as blood glucose level. In this paper, we aim to study the behavioural distinction of two groups of diabetes patient based on their published posts online. In particular, we investigate the behavioural distinction between patients who can successfully manage their blood glucose value and those who fail. We observed several distinction in terms of linguistic, textual and visual contents of published posts online. We also provide a supervised approach to predict the success of users based on his online behaviour.

# 9. REFERENCES


[1] J. Alvarez-Conrad, L. A. Zoellner, and E. B. Foa. Linguistic predictors of trauma pathology and physical health. *Applied Cognitive Psychology*, 15(7):S159–S170, 2001.

[2] R. J. Anderson, K. E. Freedland, R. E. Clouse, and P. J. Lustman. The prevalence of comorbid depression in adults with diabetes a meta-analysis. *Diabetes care*, 24(6):1069–1078, 2001.

[3] A. D. Association et al. Standards of medical care in diabetes–2012. *Diabetes care*, 35:S11, 2012.

[4] D. J. Attai, M. S. Cowher, M. Al-Hamadani, J. M. Schoger, A. C. Staley, and J. Landercasper. Twitter social media is an effective tool for breast cancer patient education and support: patient-reported outcomes by survey. *Journal of medical Internet research*, 17(7):e188, 2015.

[5] Y. Cao, F. Liu, P. Simpson, L. Antieau, A. Bennett, J. J. Cimino, J. Ely, and H. Yu. Askhermes: An online question answering system for complex clinical questions. *Journal of biomedical informatics*, 44(2):277–288, 2011.

[6] C.-H. Chang, M. Kayed, M. R. Girgis, and K. F. Shaala. A survey of web information extraction systems. *Knowledge and Data Engineering, IEEE Transactions on*, 18(10):1411–1428, 2006.

[7] D. P. Chapman, G. S. Perry, and T. W. Strine. The vital link between chronic disease and depressive disorders. *Prev Chronic Dis*, 2(1):A14, 2005.

[8] Z. Che, S. Purushotham, R. Khemani, and Y. Liu. Distilling knowledge from deep networks with applications to healthcare domain. *arXiv preprint arXiv:1512.03542*, 2015.

[9] M. A. Davis, D. L. Anthony, and S. D. Pauls. Seeking and receiving social support on facebook for surgery. *Social Science & Medicine*, 131:40–47, 2015.

[10] M. De Choudhury. Anorexia on tumblr: A characterization study. In *Proceedings of the 5th*

*International Conference on Digital Health 2015*, pages 43–50. ACM, 2015.

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

[12] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz. Predicting depression via social media. In *ICWSM*, 2013.

[13] M. De Choudhury, M. R. Morris, and R. W. White. Seeking and sharing health information online: comparing search engines and social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1365–1376. ACM, 2014.

[14] M. Dredze. How social media will change public health. *Intelligent Systems, IEEE*, 2012.

[15] B. Fogg and J. Hreha. Behavior wizard: a method for matching target behaviors with solutions. In *Persuasive technology*, pages 117–131. Springer, 2010.

[16] S. Fox and K. Purcell. Social media and health. `http://www.pewinternet.org/2010/03/24/social-media-and-health/`. Accessed: 2016-02-30.

[17] M. Franciosi, F. Pellegrini, G. De Berardis, M. Belfiglio, D. Cavaliere, B. Di Nardo, S. Greenfield, S. H. Kaplan, M. Sacco, G. Tognoni, et al. The impact of blood glucose self-monitoring on metabolic control and quality of life in type 2 diabetic patients an urgent need for better educational strategies. *Diabetes care*, 24(11):1870–1877, 2001.

[18] D. G. Garrett and B. M. Bluml. Patient self-management program for diabetes: first-year clinical, humanistic, and economic outcomes. *Journal of the American Pharmacists Association*, 45(2):130–137, 2005.

[19] N. J. Gray, J. D. Klein, P. R. Noyce, T. S. Sesselberg, and J. A. Cantrill. Health information-seeking behaviour in adolescence: the place of the internet. *Social science & medicine*, 60(7):1467–1478, 2005.

[20] J. A. Greene, N. K. Choudhry, E. Kilabuk, and W. H. Shrank. Online social networking by patients with diabetes: a qualitative evaluation of communication with facebook. *Journal of general internal medicine*, 26(3):287–292, 2011.

[21] L. Haas, M. Maryniuk, J. Beck, C. E. Cox, P. Duker, L. Edwards, E. B. Fisher, L. Hanson, D. Kent, L. Kolb, et al. National standards for diabetes self-management education and support. *Diabetes care*, 36(Supplement 1):S100–S108, 2013.

[22] C. Hawn. Take two aspirin and tweet me in the morning: how twitter, facebook, and other social media are reshaping health care. *Health affairs*, 28(2):361–368, 2009.

[23] S. Jamison-Powell, C. Linehan, L. Daley, A. Garbett, and S. Lawson. I can't get no sleep: discussing# insomnia on twitter. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1501–1510. ACM, 2012.

[24] W. Katon and M. D. Sullivan. Depression and chronic medical illness. *J Clin Psychiatry*, 51(Suppl 6):3–11, 1990.

[25] H. C. Kelman. Compliance, identification, and internalization: Three processes of attitude change. *Journal of conflict resolution*, pages 51–60, 1958.

[26] S. Klein, N. F. Sheard, X. Pi-Sunyer, A. Daly, J. Wylie-Rosett, K. Kulkarni, and N. G. Clark. Weight management through lifestyle modification for the prevention and management of type 2 diabetes: Rationale and strategies a statement of the american diabetes association, the north american association for the study of obesity, and the american society for clinical nutrition. *Diabetes care*, 27(8):2067–2073, 2004.

[27] H. Korda and Z. Itani. Harnessing social media for health promotion and behavior change. *Health promotion practice*, 14(1):15–23, 2013.

[28] M. Kumar, M. Dredze, G. Coppersmith, and M. De Choudhury. Detecting changes in suicide content manifested in social media following celebrity suicides. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*, pages 85–94. ACM, 2015.

[29] B. Liu. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167, 2012.

[30] K. Lorig, P. L. Ritter, D. D. Laurent, K. Plant, M. Green, V. B. B. Jernigan, and S. Case. Online diabetes self-management program a randomized study. *Diabetes care*, 33(6):1275–1281, 2010.

[31] D. D. Luxton, J. D. June, and J. M. Fairall. Social media and suicide: a public health perspective. *American Journal of Public Health*, 102(S2):S195–S200, 2012.

[32] D. MacLean, S. Gupta, A. Lembke, C. Manning, and J. Heer. Forum77: An analysis of an online health forum dedicated to addiction recovery. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 1511–1526. ACM, 2015.

[33] E. W. Maibach and D. Cotton. Moving people to behavior change: a staged social cognitive approach to message design. 1995.

[34] U. L. Malanda, S. D. Bot, and G. Nijpels. Self-monitoring of blood glucose in noninsulin-using type 2 diabetic patients it is time to face the evidence. *Diabetes Care*, 36(1):176–178, 2013.

[35] C. J. Murray and A. D. Lopez. Evidence-based health policy–lessons from the global burden of disease study. *Science*, 274(5288):740, 1996.

[36] K. Park, I. Weber, M. Cha, and C. Lee. Persistent sharing of fitness app status on twitter. *arXiv preprint arXiv:1510.04049*, 2015.

[37] M. J. Paul and M. Dredze. Discovering health topics in social media using topic models. *PLoS One*, 9(8):e103408, 2014.

[38] J. W. Pennebaker, M. E. Francis, and R. J. Booth. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71:2001, 2001.

[39] P. N. Robinson. Deep phenotyping for precision medicine. *Human mutation*, 33(5):777–780, 2012.

[40] D. Ruths, J. Pfeffer, et al. Social media for large studies of behavior. *Science*, 346(6213):1063–1064, 2014.

[41] H. A. Schwartz, M. Sap, M. L. Kern, J. C. Eichstaedt, A. Kapelner, M. Agrawal, E. Blanco, L. Dziurzynski, G. Park, D. STILLWELL, et al. Predicting individual well-being through the language of social media. In *Biocomputing 2016: Proceedings of the Pacific Symposium*, pages 516–527, 2016.

[42] D. A. Shoham, L. Tong, P. J. Lamberson, A. H. Auchincloss, J. Zhang, L. Dugas, J. S. Kaufman, R. S. Cooper, and A. Luke. An actor-based model of social network influence on adolescent body size, screen time, and playing sports. *PloS one*, 7(6):e39795, 2012.

[43] Z. Sun, F. Wang, and J. Hu. Linkage: An approach for comprehensive risk prediction for care management. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1145–1154. ACM, 2015.

[44] L. Tang, X. Wang, and H. Liu. Group profiling for understanding social structures. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(1):15, 2011.

[45] S. E. Taylor and J. D. Brown. Illusion and well-being: a social psychological perspective on mental health. *Psychological bulletin*, 103(2):193, 1988.

[46] M. Thelen and E. Riloff. A bootstrapping method for learning semantic lexicons using extraction pattern contexts. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 214–221. Association for Computational Linguistics, 2002.

[47] J. Vosecky, D. Hong, and V. Y. Shen. User identification across multiple social networks. In *Networked Digital Technologies, 2009. NDT'09. First International Conference on*, pages 360–365. IEEE, 2009.

[48] H. M. Wallach, I. Murray, R. Salakhutdinov, and D. Mimno. Evaluation methods for topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1105–1112. ACM, 2009.

[49] Y.-C. Wang, R. Kraut, and J. M. Levine. To stay or leave?: the relationship of emotional and informational support to commitment in online health support groups. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, pages 833–842. ACM, 2012.