Twitter User Representation Using Weakly Supervised Graph Embedding

Tunazzina Islam, Dan Goldwasser
Department of Computer Science, Purdue University, West Lafayette, Indiana 47907
islam32@purdue.edu, dgoldwas@purdue.edu

Abstract
Social media platforms provide convenient means for users to participate in multiple online activities on various contents and create fast widespread interactions. However, this rapidly growing access has also increased the diverse information, and characterizing user types to understand people’s lifestyle decisions shared in social media is challenging. In this paper, we propose a weakly supervised graph embedding based framework for understanding user types. We evaluate the user embedding learned using weak supervision over well-being related tweets from Twitter, focusing on ‘Yoga’, ‘Keto diet’. Experiments on real-world datasets demonstrate that the proposed framework outperforms the baselines for detecting user types. Finally, we illustrate data analysis on different types of users (e.g., practitioner vs. promotional) from our dataset. While we focus on lifestyle-related tweets (i.e., yoga, keto), our method for constructing user representation readily generalizes to other domains.

Introduction
Social media has been rapidly evolved to make inferences about the real world, with application to health (Dredze 2012; De Choudhury et al. 2013; Pantic 2014), well-being (Schwartz et al. 2013a), politics (O’Connor et al. 2010), marketing (Gopinath, Thomas, and Krishnamurthi 2014). Over the last decade, we have witnessed a dramatic change in social media microblogging platforms, specifically Twitter. An increasing number of people access Twitter to express opinions, engage with friends, share ideas & thoughts, and propagate various contents in their social circles.

Motivation
Twitter is a highly influential and relevant resource for understanding lifestyle choices, health, and well-being (Schwartz et al. 2016; Yang and Srinivasan 2016; Amir et al. 2017; Reece et al. 2017; Islam 2019). The information provided is often shaped by people’s underlying lifestyle choices, motivations, and interests. Besides, multiple commercial parties, practitioners, and interest groups use this platform to advance their interests and share their journey focusing on specific lifestyle decisions based on different motivations. Fig. 1 shows two different Twitter users’ profile description and tweets focusing on keto diet. Though both users susan and keto_collab tweet about the keto diet, they have different intentions. User susan (yellow box) is a ‘keto enthusiast’ tweeting about her keto journey, while keto_collab (green box) focuses on collecting information from various keto channels and promoting keto recipes and ketogenic diet daily. In addition to the tweets’ contents, the profile description of susan indicates that she is a practitioner. On the other hand, the profile description and tweets of keto_collab indicate that it is a promotional account (Fig. 1). Our goal is to automatically classify user types from well-being related tweets and analyze their textual content.

To demonstrate our proposed method to characterize user types, we consider two lifestyle-related activities: Yoga – a popular multi-faceted activity focusing on the body-mind-spirit connection (Goyeche 1979) having benefits of alleviating symptoms of anxiety and depression as well as promoting good physical fitness (Khalsa 2004; Yurtkuran, Alp, and Dilek 2007; Smith and Pukall 2009; Ross and Thomas 2010) and Keto diet – a low-carbohydrate, high-fat, adequate-protein diet helping weight loss (Johnstone et al. 2008), controlling type − 2 diabetes (McKenzie et al. 2017) as well as therapeutic potential in pathological conditions, such as PCOS, acne, neurological diseases, cancer, and the amelioration of respiratory and cardiovascular disease risk factors (Paoli et al. 2013). Despite the current popularity of yoga and keto, there is little research on analyzing users’ lifestyle decision in social media. However, understanding user dynamics on social media is critical to analyze their lifestyle choice and motivation.

Challenges
First, social media users have massive and diverse information regarding topics, content, demographics. It is hard to obtain large-scale annotated data for training a machine learning models on such sophisticated content. Second, tweets are short and often ambiguous. So, a simple pattern-based analysis on text (tweets) using specific keywords is often inadequate for capturing relevant information. Third, user representation requires multiview formulation of data. A system solely relying on tweets’ contents misses the rich auxiliary information from the available sources. In addition to the tweet, user’s profile description and network are helpful. In this work, we suggest a graph embedding based approach.

Copyright © 2022, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.
Figure 1: Two different Twitter users’ profile sharing their interests on keto diet. Here, Des represents users’ profile description and T1, T2 are users’ tweets.

@susan
Des: Boy mom, wife, Engineer, Zumba Instructor, Keto Enthusiast.
T1: #titleaders my Keto Pancakes recipe: 4 eggs, 4 oz cream cheese, 1/2 cup almond flour, fresh blueberries Pancakes. #ketolife
T2: Almost year 4 on Keto and finally found a cereal substitute #ketodiet #granola #HealthyEating

@keto_collab
Des: We are Ketogenic Information Collaborator. We collect information from Various Keto channels and Tweet it out for you.
T1: Keto Frosted Flakes Cereal Recipe - Low Carb "Corn Flakes Alternative" https://myketokitchen.com/keto-recipes/

designed to address the above challenges for generating user representation by leveraging weak supervision from profile description.

Existing Work
Prior works infer many latent user characteristics like personality (Golbeck et al. 2011; Kosinski, Stillwell, and Graepel 2013; Schwartz et al. 2013b; Lynn, Balasubramanian, and Schwartz 2020), emotions (Wang and Pal 2015), happiness (Islam and Goldwasser 2020), mental health (Amir et al. 2017), mental disorders (De Choudhury et al. 2013; Reece et al. 2017) by analyzing the information in a user’s social media account. Several works have been done on multiview formulation of data in social media analysis research. (Islam and Goldwasser 2020) investigated relationship between practicing yoga and being happy by incorporating textual and temporal information of users using Granger causality. (Mishra, Yannakoudakis, and Shutova 2018; Mishra et al. 2019) exploited user’s community information along with textual features for detecting abusive instances. (Ribeiro et al. 2018) characterized hateful users using content as well as user’s activity and connections. (Miura et al. 2017; Ebrahimi et al. 2018; Huang and Carley 2019) used a joint model incorporating different types of available information, including tweet text, user network, and metadata for geolocating Twitter users. These works are typically studied as a supervised learning task. Previous works aiming to understand Twitter users’ types by combining their tweets, metadata, and social information (Del Tredici et al. 2019; Islam and Goldwasser 2021b,a) rely on large amounts of labeled instances to train supervised models. Such sizeable labeled training data is difficult to obtain. Semi-supervised methods can reduce dependence on labeled texts. Graph based semi-supervised algorithms achieved considerable attention over the years (Belkin, Niyogi, and Sindhwani 2006; Subramanya and Bilmes 2008; Talukdar et al. 2008). (Sindhwani and Melville 2008) used a polarity lexicon combined with label propagation (Zhu and Ghahramani 2002) for sentiment analysis. Several have used label propagation for polarity classification (Blair-Goldensohn et al. 2008; Brody and Elhadad 2010), community detection (Jokar and Mosleh 2019).

Our challenge is to construct a user representation in a weakly supervised way relevant for characterizing nuanced activity and lifestyle-specific properties.

Contributions
In this paper, our goal is to explore the approach driven by the principle of social homophily (McPherson, Smith-Lovin, and Cook 2001), indicating individuals’ tendency to form social ties with others who have common interests. (Yang and Eisenstein 2017) used this phenomenon to overcome language variation in sentiment analysis. Graph-based method such as label propagation (Zhu and Ghahramani 2002; Talukdar and Crammer 2009; Speriosu et al. 2011) provides a representation by exploiting such relationships to improve classification, often while requiring less supervision than with standard classification. In our settings, we follow the observation that the users’ lifestyle choices expressed in the
Figure 2: Information graph capturing relations among users, descriptions, user types.

Text will be reflected in users’ behavior engaging with it. The main insight of our work is that given two users with very similar behaviors and similar activity, their profile descriptions can be aligned. We define an objective that does the weak supervision - mapping profile descriptions to labels (i.e., user types) based on specific keywords.

Fig. 2 represents the relationship between users, descriptions, user types. To capture more hidden relationships that characterize user types through information graph, we suggest inference function to apply iteratively. Fig. 2 shows that users $u_1$ and $u_3$ have weak labels that we obtain from their profile descriptions $d_1$ and $d_3$ respectively by mapping with labels using specific keywords. Profile description of $u_2$ doesn’t have label initially. Inference function infers the label of user $u_2$ and $u_2$’s description, $d_2$ as ‘practitioner’ because $u_1$ and $u_2$ have similar tweets (Fig. 2).

From a technical viewpoint, we define our user type detection as a reasoning problem over an information graph that creates distributed representations of nodes contextualized by the graph structure, allowing us to transfer information from observed nodes to unknown nodes (Step 1). Step 2 – we use inference function built on similarity metric defined by the learned graph embedding to increase the number of edges connecting the two node types. We suggest an Expectation–maximization (EM) approach for completing these two steps. Our contributions can be summarized as follows:

1. We formulate a novel problem of exploiting weak supervision for user type detection from social media.
2. We suggest a graph embedding based EM-style approach for learning and reasoning to construct like-minded users incrementally.
3. We describe how to generate weak labels from user’s profile description along with quantitative quality assessment.
4. We conduct extensive experiments on real-world datasets to demonstrate the effectiveness of the proposed weakly supervised graph embedding method to detect user type over the baselines.

Model

In this section, we first describe how to create the information graph and learn the information. Next, we discuss about inference function. Then, we discuss an iterative EM-style approach that continually improves the user representation learned by the graph.

Information Graph Creation

We represent users’ activity on social media as an information graph, connecting profile description to user representing by their tweets. In this process, we have following nodes: users representing by their tweets, profile descriptions, user types (practitioner, promotional). We have following observed edges: profile description-to-user type, user-to-user type, profile description-to-user, user-user. As most users’ descriptions are not associated with the labels, our technical challenge is to infer the unknown label of the descriptions. To address this challenge, we learn a graph embedding to maximize the similarity between neighboring nodes (Perozzi, Al-Rfou, and Skiena 2014; Tang et al. 2015; Grover and Leskovec 2016).
Information Graph Embedding

We define the notion of information graph shown in Fig. 2. Let Graph, \( G = (V, E) \), where \( V \) consists of the following nodes: (1) \( D = \{d_1, d_2, \ldots, d_m\} \) represents user’s profile description, (2) \( U = \{u_1, u_2, \ldots, u_n\} \) are the Twitter users representing by their tweets, (3) \( UT = \{u_1t, u_2t, \ldots, u_nt\} \) represents user type (label) which is binary (i.e., practitioner and promotional) in our case. \( E \) contains following edges: (1) each Twitter user must have profile description so \( D \) connects with \( U \), (2) profile description has connection with user type based on specific keywords resulting a connection between \( D \) and \( UT \), (3) user \( U \) connects with user type \( UT \) if the user’s description is already connected to user type, (4) \( U \) and \( U \) via follow link. For follow link, we consider those users from our dataset if they are retweeted and/or @-mentioned (Rahimi et al. 2015) in other users’ (from our data) tweets. An edge is created between two users if either user mentioned the other from the data. We embed the following instances in a common embedding space - (a) users, (b) profile descriptions, (c) user types. We maximize the similarity between two instances in the embedding space if – (1) profile description has a type. (2) a user has a type.

Let define the embedding function, \( \phi \) that maps the graph’s nodes to vectors \( R^d \). We train this by following a negative sampling approach. As we construct positive examples by associating each node \( u \) with all of its neighbors, \( (u, xp) \) is a positive pair where \( xp \) is a positive example. For each positive pair, we sample 5 negative examples \( (x^n) \) such that way that \( u \) and \( x^n \) do not share an edge. Our goal is to maximize the similarity of a node embedding with a positive example and minimize the similarity with a negative example. We call a user type a positive example for a user if the user belongs to that type. Otherwise, the type is called a negative example. The embedding loss is designed to place \( u \) closer to \( xp \) than \( x^n \). We define the loss function as follows:

\[
E_t = l(\text{sim}(\phi(u), \phi(xp)), \text{sim}(\phi(u), \phi(x^n)))
\]

(1)

where, \( E_t \) defines the embedding loss for specific objective function \( t \) (i.e., user to user type). Similarity function, \( \text{sim}() \) is the dot product and \( l() \) is the cross-entropy loss.

\[
l(p, n) = -\log\left(\frac{e^{\text{sim}(\phi(u), \phi(xp))}}{e^{\text{sim}(\phi(u), \phi(xp))} + e^{\text{sim}(\phi(u), \phi(x^n))}}\right)
\]

(2)

We minimize the summed loss \( \sum_{t \in T} \lambda_t E_t \) where \( T \) is the set of all objective functions and the weight, \( \lambda_t \) associates with each objective function.

We obtain user embeddings by running a Bi-LSTM (Schuster and Paliwal 1997) over the Glove (Pennington, Socher, and Manning 2014) word embeddings (300d) of the words of the user’s tweets. Concatenating the hidden states of the two opposite directional LSTMs, we get representation over one time-stamp and average the representations of all time-stamps to obtain a single representation of the tweets. We train this Bi-LSTM jointly with embedding learning.

Inference Function

After learning the information from graph embedding, which captures the observed relations, we can obtain unobserved relations among nodes in the graph. Based on the similarity captured between non-connected nodes by graph embedding, we create new edges between them to use this knowledge in the future. We do this by defining an inference function that creates the edges based on information graph inferences. For example, we have users whose profile descriptions are not covered by specific keywords used to create a weak label. But those users may have similar tweet contents that match with the tweets of a labeled user. The graph embedding step learn this commonality between these users and represent them with similar embeddings. Using inference function, we directly connect the user and label with an edge as well as the user’s description and label with an edge based on the user node similarity in the embedding space.

For the inference function, we make edge connections based on the node representations learned by computing similarity scores between all pairs of nodes (using the node embedding) and connecting the nodes with the top \( k \) scores.

EM-style Learning Approach

In this section, we describe overall EM-style graph learning framework that continually as follows:

Step 1: Learn node embedding. In this step, we learn the information graph embedding using the framework described in subsection named Information Graph Embedding to obtain initial graph representation.

Step 2: Infer unlabeled users. We apply inference function (see details in subsection named Inference Function) built on similarity metric based on the learned information graph representation.

Step 3: Stopping criterion. At each iteration, after Step 2, we check the predicted labels of all users. Our model converges when the change in predicted labels is less than a threshold between two consecutive iterations.

Dataset Details

To evaluate our model’s ability to characterize user type, we use two lifestyle choice related data, i.e., yoga and keto diet from Twitter. We download tweets using Tweepy\(^1\) by Twitter streaming API sub-sequentially from May to November, 2019. For yoga, we collect 419608 tweets related to yoga containing popular keywords: ‘yoga’, ‘yogi’, ‘yogalife’, ‘yogalove’, ‘yogainspiration’, ‘yogachallenge’, ‘yogaeverywhere’, ‘yogaeveryday’, ‘yogadaily’, ‘yogaeverydaymerry’, ‘yogapractice’, ‘yogapose’, ‘yogalover’, ‘yogajourney’. There are 297350 different users among them 15168 users have at least yoga-related tweets in their timelines. We have 35392177 timeline tweets in total. We discard those users who do not have profile description. So we have finally 13301 yoga users.


\( ^1\)https://www.tweepy.org/
least two keto-related tweets in their timelines having total 39250716 timeline tweets. After discarding the users not having profile description, we have 14320 keto users finally.

Holdout Data
For testing purpose, we manually annotate 786 yoga users and 908 keto users using binary label ‘practitioner’, ‘promotional’. For each user, we check both their profile description and timeline tweets (only yoga/keto-related). At first, we look at the user profile description for user type, whether they explicitly mention practicing a specific lifestyle (e.g., yogi, yogini, keto enthusiast, ketosis); then, we look into the timeline tweets of that user. If the user tweets about the first-hand experience of practicing yoga (e.g., love doing yoga during weekend morning/keto diet (e.g., lost 7lb in 3rd week of keto diet)), we annotate them as a ‘practitioner’. After looking at the description and tweets, if we observe that they are promoting a gym/studio (e.g., offering online yoga classes), online shop (e.g., selling yoga pants), app (e.g., keto diet recipe), restaurant (e.g., serving keto meal) etc., rather than sharing their first-hand experience about a particular lifestyle, we annotate the user as a ‘promotional’ user.

Two graduate students from Computer Science department manually annotate a subset of tweets (10%) for calculating inter-annotator agreement. This subset has an inter-annotator agreement of 64.7%, which is substantial agreement using Cohen’s Kappa coefficient (Cohen 1960). In case of a disagreement, we resolve it by discussion.

Information Graph
In our information graph, we have following nodes: users representing by their tweets, profile descriptions, user types and following observed edges: profile description-to-user type, user-to-user type, profile description-to-user, user-user. For yoga data, initially we have 33784 nodes with 8730 observed edges. After EM-style approach, we have total 19927 edges in our information graph for yoga. In our keto data, we have 28583 nodes and 33784 observed edges initially. After EM-style approach, we notice that our keto information graph contains total 145944 edges.

Constructing Weak Labels
In this section, we describe how to generate weak labels from users’ profile descriptions that can be incorporated as weak sources in our model. We further evaluate the quality of the weak labels.

Generating Weak Labels
Keyword based knowledge extraction from profile description. In some cases, the user’s profile description contains specific keywords indicating a particular user type (i.e., practitioner, promotional). We use those keywords to extract the label and utilize them as weak-supervision for our model. However, not to make our weak label too noisy, we do not assume that people with ‘yoga’ or ‘keto’ mentioned in their profile description would be ‘practitioner’ because other keywords that appear with them would be the indicator of being ‘promotional’. Also, the exact keywords might appear in both user types. So we need to check other keywords appearing with them as separating criteria for ‘practitioner’ and ‘promotional’. To handle these cases, we form the following rules based on more related keywords:

For yoga, if user description contains following words (‘yoga’ or ‘yogi’ or ‘yogini’ or ‘guru’ or ‘spirituality’ or ‘asana’ or ‘meditation’ or ‘mindfulness’ or ‘fitness’ or ‘health’) and not (‘studio’, ‘gym’, ‘magazine’, ‘daily’, ‘channel’, ‘clothing’, ‘subscribe’, ‘training’, ‘shop’, ‘sale’, ‘discount’, ‘package’, ‘free’, ‘ship’, ‘retreat’, ‘resort’, ‘design’, ‘jewellery’, ‘handmade’, ‘business’, ‘spa’, ‘hotel’, ‘restaurant’, ‘activewear’, ‘beachwear’, ‘sportswear’, ‘festival’, ‘app’, ‘website’, ‘community’, ‘organizer’, ‘center’, ‘donate’, ‘support’, ‘fund’, ‘product’, ‘review’, ‘trend’, ‘healthcare’), we consider them yoga practitioner. If profile description has following words (‘gym’ or ‘studio’ or ‘magazine’ or ‘daily’ or ‘channel’ or ‘clothing’ or ‘subscribe’ or ‘training’ or ‘shop’ or ‘sale’ or ‘discount’ or ‘package’ or ‘free’ or ‘ship’ or ‘retreat’ or ‘resort’ or ‘design’ or ‘jewellery’ or ‘handmade’ or ‘business’ or ‘spa’ or ‘hotel’ or ‘restaurant’ or ‘activewear’ or ‘beachwear’ or ‘sportswear’ or ‘festival’ or ‘app’ or ‘website’ or ‘community’ or ‘organizer’ or ‘center’ or ‘donate’ or ‘support’ or ‘fund’ or ‘product’ or ‘review’ or ‘trend’ or ‘healthcare’), we consider them promotional user. After applying keyword based knowledge extraction rule, we have 2104 weak labels for yoga data.


Quality of Weak Labeling
To assess the weak label quality, we consider the data with both weak label and ground-truth label. For yoga data, we have 451 users and for keto, we have 56 users having both weak and true label. We compare the weak labels with the ground truth labels. The accuracy and macro-avg F1 score of the weak label are 0.79 and 0.78 respectively for yoga data. For keto data, the accuracy and macro-avg F1 score of the weak label are 0.86 and 0.67 correspondingly. We observe that the accuracy and macro-avg F1 score of the weak label both for yoga and keto data are significantly better than random (0.5) for binary classification, indicating that our weak labeling approach has acceptable quality.

Experiments
In this section, we present the experiments to evaluate the effectiveness of our model, baselines, hyperparameter tun-
<table>
<thead>
<tr>
<th>Model</th>
<th>Yoga</th>
<th>Keto</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy</td>
<td>Macro-avg F1</td>
</tr>
<tr>
<td>LSTM_Glove</td>
<td>0.51</td>
<td>0.45</td>
</tr>
<tr>
<td>Fine-tuned BERT</td>
<td>0.47</td>
<td>0.47</td>
</tr>
<tr>
<td>Label propagation</td>
<td>0.78</td>
<td>0.75</td>
</tr>
<tr>
<td>EM-style approach</td>
<td>0.78</td>
<td>0.76</td>
</tr>
</tbody>
</table>

Table 1: The first two rows are supervised baselines. The last two rows show the user type detection results for weak supervision settings.

<table>
<thead>
<tr>
<th>Model</th>
<th>Yoga</th>
<th>Keto</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy</td>
<td>Macro-avg F1</td>
</tr>
<tr>
<td>Label propagation (des)</td>
<td>0.721</td>
<td>0.711</td>
</tr>
<tr>
<td>EM-style approach (des)</td>
<td>0.781</td>
<td>0.761</td>
</tr>
<tr>
<td>Label propagation (net)</td>
<td>0.573</td>
<td>0.572</td>
</tr>
<tr>
<td>EM-style approach (net)</td>
<td>0.670</td>
<td>0.657</td>
</tr>
<tr>
<td>Label propagation (des + net)</td>
<td>0.781</td>
<td>0.753</td>
</tr>
<tr>
<td>EM-style approach (des + net)</td>
<td>0.782</td>
<td>0.763</td>
</tr>
</tbody>
</table>

des : profile description  
net : user network  
des + net : both profile description and user network

Table 2: Ablation study.

Experimental Settings

We use accuracy and macro-average F1 score as the evaluation metrics. We show two versions of our model (weakly supervised) — (1) initial graph embedding without iterating multiple times (Label propagation), (2) EM-style approach. Both of them are trained on weakly labeled data and evaluated with the performance on holdout data (manually annotated ground-truth discussed in section named Dataset Details). Both of our weakly supervised models are repeated for 3 times and we report the average performance (3rd and 4th rows of Table 1). We compare our model with two supervised learning baselines. Table 1 shows that our EM-style graph embedding model achieves the highest performance both in accuracy (yoga = 78.2%, keto = 72.2%) and macro-avg F1 score (yoga = 76.3%, keto = 64.2%) than the supervised baselines which answer the EQ1.

Baselines. For supervised baselines, we train a model using the weak labels assigned by the keywords and make predictions about the users that do not match any keyword. From the weakly labeled training data, we randomly choose 20% data as validation set. Our first supervised baseline is referred to as LSTM_Glove, where we use 300d Glove word embeddings to obtain the embedding of user’s tweets and forward these embeddings to LSTM (Hochreiter and Schmidhuber 1997). We use cross-entropy loss. We fine-tune the pre-trained BERT (base-uncased) (Devlin et al. 2019) model with user’s tweets as our second supervised baseline. For both baselines, we use 3-fold cross-validation and the average is reported in Table 1.

Hyperparameter Details. We run our proposed model for characterizing user type on yoga and keto data. For both of them, we run the embedding learning (described in subsection named Information Graph Embedding) at most 100 epochs or stop the learning if the embedding learning loss does not decrease for 10 consecutive epochs. For inference function, we pick the top 20 most confident predictions based on majority voting and treat them as labeled users. For stopping criterion, we set the threshold to 10%. We end up running inference function for 2 iterations for yoga and 3 iterations for keto. We use optimizer= Adam (Kingma and Ba 2014), learning rate= 0.001, batch size = 16. The single layer Bi-LSTM takes 300d Glove word embeddings as inputs and maps to a 150d hidden layer. We initialize the embeddings of all of the other instances randomly in 300d space. We initialize the weight associates with each objective function, $\lambda_t = 1$, for all.

For LSTM_Glove baseline, the single layer LSTM takes 300d Glove word embeddings as inputs and maps to a 150d hidden layer with optimizer= Adam, learning rate= 0.01, batch size = 16, epochs = 20. For BERT fine-tuning, to encode our texts, for padding or truncating we decide maximum sentence length = 500. We use batch size = 32, learning rate
observed edges. After the EM-style approach, we have total
5811 edges in the information graph of keto. For keto data,
we have 28583 nodes and 855 observed edges initially. After
the EM-style approach, we notice that our information graph
contains total 14313 edges for keto. Table 2 (2nd row) shows
that EM-style graph embedding model with description only
achieves 78.1% accuracy for yoga and 66.4% for keto as well
as 76.1% macro-avg F1 score for yoga and 63.5% for keto.

Then, we train both Label propagation and EM-style ap-
proach with user network only (3rd and 4th rows of Table
2). Information graph contains the following nodes: users representing by their tweets, profile descriptions, user types (practitioner, promotional). We have the following observed edges: profile description-to-user type, user-to-user type, profile description-to-user. For yoga data, initially, we have 26482 nodes with 2104 ob-
served edges. After the EM-style approach, we have total 13301 edges in the information graph of yoga. For keto data, we have 28583 nodes and 855 observed edges initially. After the EM-style approach, we observe the words
‘yoga’ and ‘keto’ because of the apparent high occurrences.
Looking at yoga practitioners’ topics, we notice that they
tweet mostly about their yoga practice, yoga pose, yoga class,
yoga practice time (Topic 0 in Fig. 4a). On the other hand,
topic 1 of practitioner mostly focus on practitioners’ motiv-
ation of doing yoga, i.e., health and fitness benefit, spiritual
connection, meditation. Topics from promotional yoga users
mostly focus on joining yoga class, gym, fitness model, sport,
yoga pant (Fig. 4b). For keto practitioners, the tweets’ topics
are mostly related to ketogenic diet, low carb high fat food
(lchf), ketonemia (presence of abnormally high concentra-
tion of ketone bodies in the blood), budget friendly ketodiet
(Fig. 4c). Promotional keto users tweet mostly about keto
recipe, paleo diet, gluten-free meal, supplements, anti-aging
medicine (Fig. 4d).

We perform text analysis of user’s profile description to
understand what kind of words users use in their description
which can distinguish their types based on their lifestyle. We
create wordcloud with the most frequent words (Fig. 5) from
yoga and keto users’ profile description, filtering out the word
‘yoga’ and ‘keto’ respectively, because of the apparent high
occurrences. We observe the words {‘teacher’, ‘instructor’,
in yoga practitioners’ descriptions (Fig. 5a). Promotional users
have the following words in the description {‘studio’,
‘event’, ‘business’} (Fig. 5b). Fig. 5c shows the wordcloud
of the profile description of keto practitioner having words
‘family’}. Promotional users’ description contains {‘free’,
‘food’, ‘recipe’, ‘health’, ‘share’, ‘bloggier’} words. We no-
tice that most users use {‘love’, ‘life’, ‘live’, ‘health’} these

Does Multiview Information Help?
Social media user representation requires multiview formu-
lation of data. To understand the contribution of the model
components, we perform an ablation study. Table 2 shows
the results of our ablation study both for yoga and keto data.

We train both initial graph embedding (Label propagation)
and EM-style approach with user description only (1st and
2nd rows of Table 2). In this case, our information graph con-
tains the following nodes: users representing by their tweets,
profile descriptions, user types (practitioner, promotional). We
have the following observed edges: profile description-
to-user type, user-to-user type, profile description-to-user. For yoga data, initially, we have 26482 nodes with 2104 ob-
served edges. After the EM-style approach, we have total 13301 edges in the information graph of yoga. For keto data,
we have 28583 nodes and 855 observed edges initially. After the EM-style approach, we notice that our information graph
contains total 14313 edges for keto. Table 2 (2nd row) shows
that EM-style graph embedding model with description only
achieves 78.1% accuracy for yoga and 66.4% for keto as well
as 76.1% macro-avg F1 score for yoga and 63.5% for keto.

Embedding Analysis
Our model’s embeddings are effective to characterize user
types. Before our model is trained, user embeddings are ran-
dom. To answer the EQ3, we project the embeddings to a 2D
space using t-SNE (Van der Maaten and Hinton 2008) and
use different colors to represent labels (Fig. 3). We plot the
user embeddings after label propagation (Fig. 3a and Fig. 3c)
and EM-style approach (Fig. 3b and Fig. 3d) based on ground
truth data. The embeddings show that EM-style approach has
better user representation both for yoga (Fig. 3b) and keto
(Fig. 3d) dataset.

User Type Analysis
In this section, we perform analysis on predicted user types
from our data. To understand the topic of the user’s tweets,
we run LDA (Blei, Ng, and Jordan 2003) based topic model-
ing over the tweets. Fig. 4 shows the wordclouds of top 10
keywords in each topic (we show 2 topics here) from prac-
titioner and promotional users’ tweets both from yoga and
keto dataset. For visualization purpose, we filter out the word
‘yoga’ and ‘keto’ because of the apparent high occurrences.
Looking at yoga practitioners’ topics, we notice that they
tweet mostly about their yoga practice, yoga pose, yoga class,
yoga practice time (Topic 0 in Fig. 4a). On the other hand,
topic 1 of practitioner mostly focus on practitioners’ motiv-
ation of doing yoga, i.e., health and fitness benefit, spiritual
connection, meditation. Topics from promotional yoga users
mostly focus on joining yoga class, gym, fitness model, sport,
yoga pant (Fig. 4b). For keto practitioners, the tweets’ topics
are mostly related to ketogenic diet, low carb high fat food
(lchf), ketonemia (presence of abnormally high concentra-
tion of ketone bodies in the blood), budget friendly ketodiet
(Fig. 4c). Promotional keto users tweet mostly about keto
recipe, paleo diet, gluten-free meal, supplements, anti-aging
medicine (Fig. 4d).
Figure 3: User embeddings after label propagation and EM-style approach both for yoga and keto data in t-SNE visualization.
Users’ Sentiment Analysis

We analyze the sentiment of practitioners’ tweets both for yoga and keto using VADER (Hutto and Gilbert 2014) model which is used for text sentiment analysis that is sensitive to both polarity (positive/negative) and intensity (strength) of emotion.

For this analysis, we randomly select 350 yoga practitioners and 150 keto practitioners and choose their timeline tweets related to yoga and keto diet respectively. After running VADER sentiment analyzer, we obtain 9147 positive sentiment and 853 negative sentiment tweets for yoga practitioners. For keto practitioners, we have 7931 positive and 2069 negative sentiment tweets.

To understand the topic of the practitioners’ tweets which have positive sentiment, we run LDA based topic modeling over the tweets. Fig. 6 shows the wordclouds of top 10 keywords in each topic (we show 2 topics here) from practitioners’ tweets which have positive sentiment both from yoga (Fig. 6a) and keto (Fig. 6b) dataset.

For yoga practitioners positive sentiment tweets, Topic 0 (Fig. 6a) is mostly related to ‘vinyasa’ and ‘hatha’ yoga. Vinyasa is a style of yoga characterized by stringing postures together so that we can move from one to another, seamlessly, using breath. It is commonly referred to as ‘flow’ yoga. Hatha yoga allows for more stretching. Topic 1 is more about joining yoga class, studio. Looking at keto practitioners’ (having positive sentiment) topics, we notice that they tweet mostly about the keto/lowcarb recipe, sugar free diet (Topic 0 in Fig. 6b). On the other hand, topic 1 of practitioners mostly focus on practitioners’ positive sentiment about weight loss, diet, meal plan.

Conclusion and Future Work

In this paper, we propose a weakly supervised graph embedding based EM-style framework to characterize user types in social media. To demonstrate our model’s effectiveness, we
perform extensive experiments on real-world datasets focusing on ‘yoga’ & ‘keto diet’ and our model outperforms the baselines. We perform data analysis on different user types from our data to show the topics from users’ tweets and frequently used words in users’ descriptions. Using our model with minimal tweaking in keywords selection for generating weak labels, we can distinguish any user types though they are not from the yoga/keto lifestyle communities. Our work can lead to new discussions on the analysis of health users’ account types. While we focus specifically on lifestyle-related tweets, our approach is a general framework that can be adapted to other corpora. In the future, we aim to expand our work to detect communities based on different lifestyle decisions and understand their motivations.

Our code and the data are available here.3

Ethics Statement

To the best of our knowledge no code of ethics has been violated throughout the annotations and experiments done in this paper. We use illustrative examples to show the Twitter user profile descriptions, tweets in Fig. 1. As our dataset is comprised of tweets, we will share the original data based upon request for research purpose only.

Acknowledgments

We gratefully acknowledge Md Masudur Rahman and the anonymous reviewers for their insightful comments. We further thank our annotators. This work was partially supported by an NSF CAREER award IIS-2048001.

References


De Choudhury, M.; Gamon, M.; Counts, S.; and Horvitz, E. 2013. Predicting depression via social media. In *Seventh international AAAI conference on weblogs and social media*.


