

# Twitter User Representation Using Weakly Supervised Graph Embedding

**Tunazzina Islam, Dan Goldwasser**

Department of Computer Science

Purdue University, West Lafayette, IN

ICWSM 2022

*Date: June 6-9, 2022*

# How do people talk about Life-Style & Well-Being?



**@susan**

Description : Boy mom, wife, Engineer, Zumba Instructor, Keto Enthusiast.

Tweet1: #fitleaders my Keto Pancakes recipe: 4 eggs, 4 oz cream cheese, 1/2 cup almond flour, fresh blueberries Pancakes. #ketolife

Tweet2 : Almost year 4 on Keto and finally found a cereal substitute  
#ketodiet #granola #HealthyEating



**@keto\_collab**

Description: We are Ketogenic Information Collaborator. We collect information from Various Keto channels and Tweet it out for you.





Tweet1: Keto Frosted Flakes Cereal Recipe - Low Carb "Corn Flakes Alternative" <https://myketokitchen.com/keto-recipes/>

Tweet2: The latest The Ketogenic diet Daily! <https://paper.li/KetoDietDaily>





# Our Goal

- Formulate a novel problem of exploiting weak supervision for characterizing users in social media.
- Suggest a graph embedding based Expectation–maximization (EM)-style approach.
- Conduct extensive experiments on real-world datasets to demonstrate the effectiveness of the model.

# Roadmap

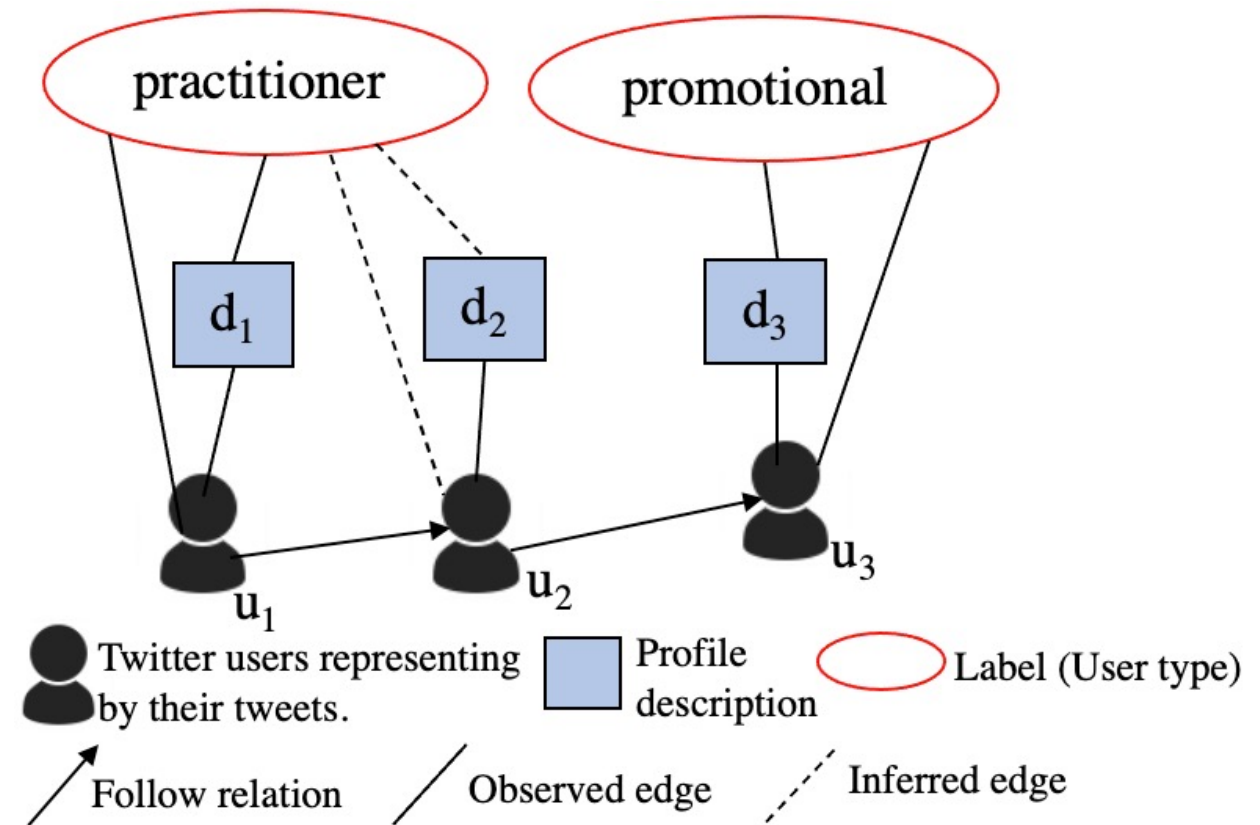
	Brief Introduction to Graph Embedding Model
	Dataset Collection and Annotation
	Automatic User Characterization
	User Type Analysis

# Roadmap

	Brief Introduction to Graph Embedding Model
	Dataset Collection and Annotation
	Automatic User Characterization
	User Type Analysis

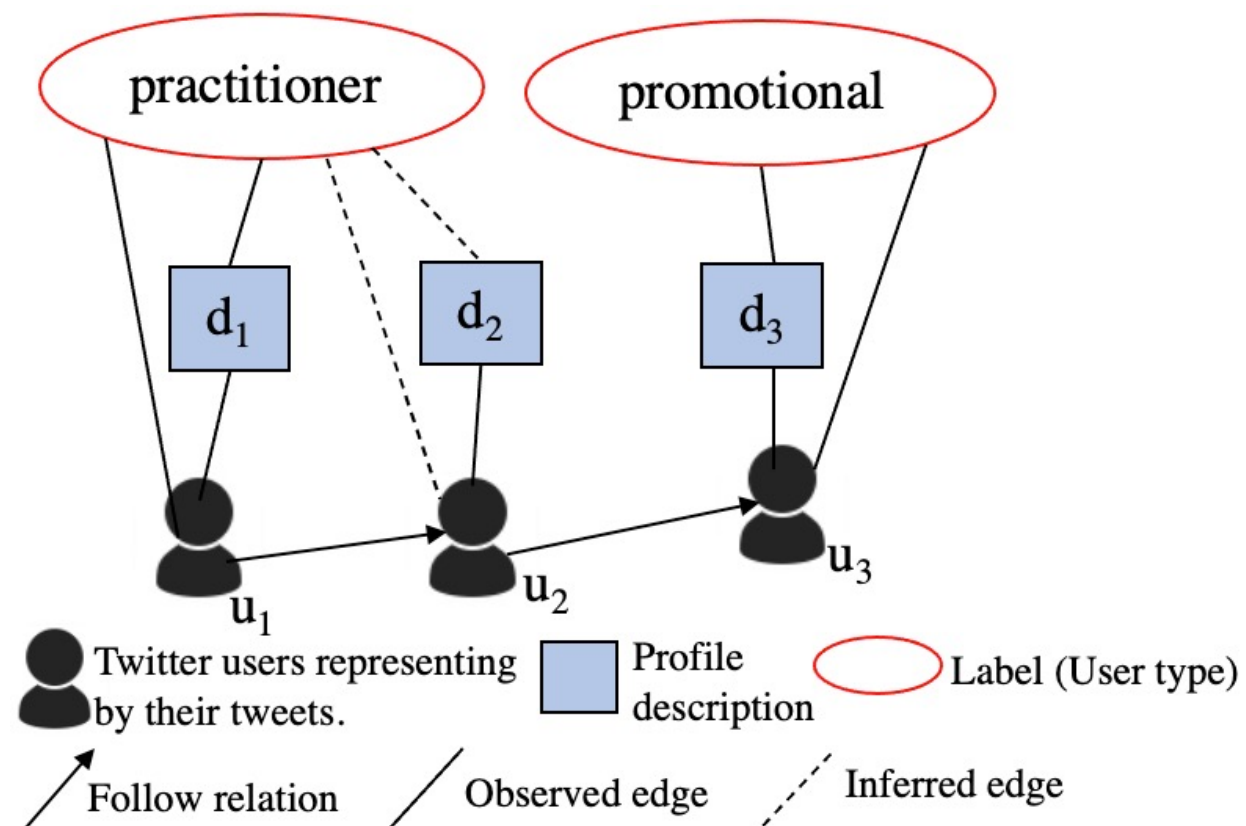
# Information Graph Creation

- Nodes:
  - users representing by tweets
  - profile description
  - user type
- Edges:
  - profile description-to-user type
  - user-to-user type
  - profile description-to-user
  - user-user



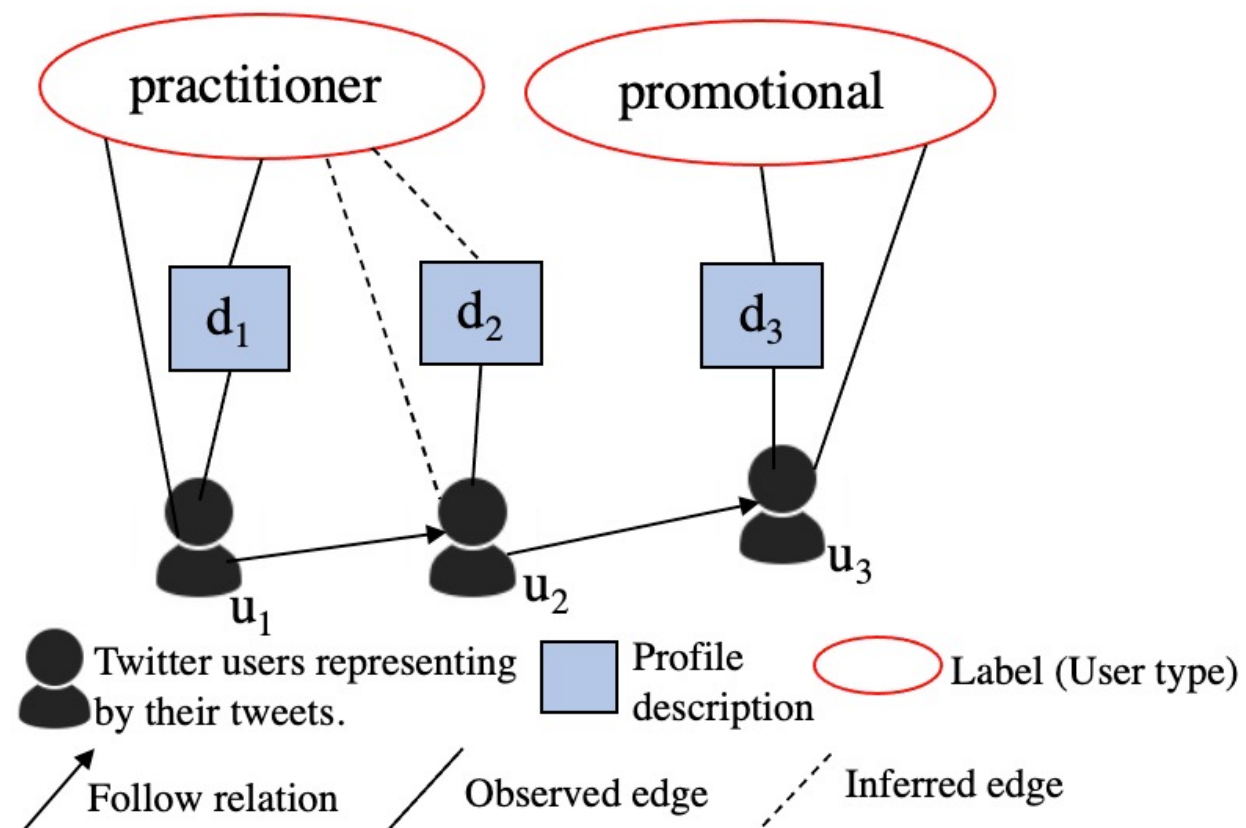
# Information Graph Embedding

- Embed nodes in a common embedding space.
- Maximize similarity between two instances in the embedding space if –
  1. profile description has a type,
  2. a user has a type.
- Train embedding following a negative sampling approach.



# Inference Function

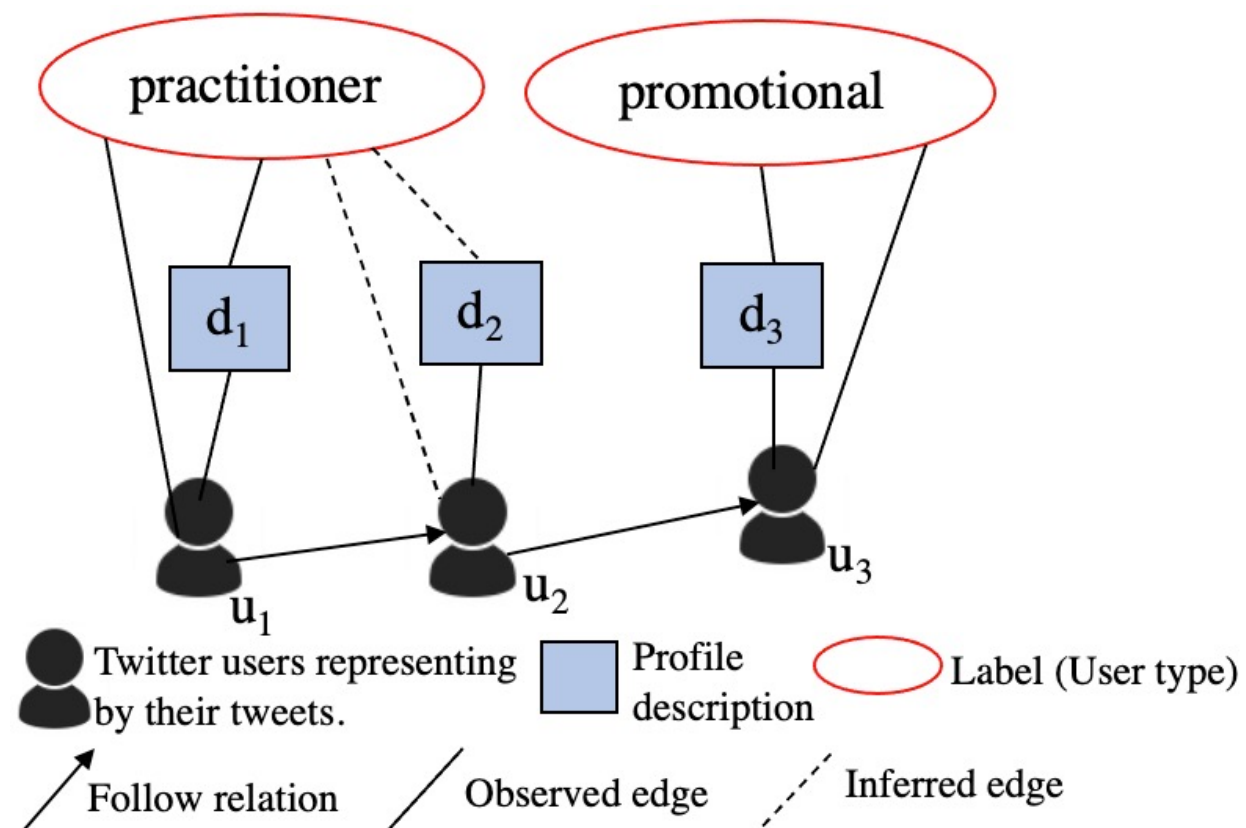
- Edge connections based on the learned node representations.
- Connecting the nodes with the top k scores.









# EM-style Learning Approach

- **Step 1:** Learn information graph embedding.
- **Step 2:** Apply inference function to infer unlabeled users.
- **Step 3:** Stopping criterion.
  - At each iteration, after Step 2, check the model convergence.







# Roadmap

	Brief Introduction to Graph Embedding Model
	Dataset Collection and Annotation
	Automatic User Characterization
	User Type Analysis

# Dataset

- **13k yoga users** and **14k keto users** from May-November 2019 from Twitter.
- Holdout Data Annotation :
  - Manually annotated **786 yoga users** and **908 keto users** using binary label ‘practitioner’, ‘promotional’.
  - 1 annotator, with annotation instruction and examples provided.
  - To calculate % agreement, 2 graduate students annotate a subset of tweets having inter-annotator agreement **65%** (substantial agreement).
- Constructing Weak Labels
  - Keyword based knowledge extraction from profile description.
- Quality of Weak Labeling:
  - 451 yoga users and 56 keto users have both weak and ground truth label
  - Yoga: accuracy **79%**, macro-avg F1 score **78%**
  - Keto: accuracy **86%**, macro-avg F1 score **67%**

# Roadmap

	Brief Introduction to Graph Embedding Model
	Dataset Collection and Annotation
	Automatic User Characterization
	User Type Analysis

# Models

- Our Model: **EM-Style Approach**
- Baseline Models:
  - Weakly Supervised Baseline:
    - Label Propagation
  - Supervised Baseline:
    - LSTM\_Glove
    - Fine-tuned BERT

Model	Yoga		Keto	
	Accuracy	Macro-avg F1	Accuracy	Macro-avg F1
LSTM_Glove	0.51	0.45	0.72	0.43
Fine-tuned BERT	0.47	0.47	0.72	0.42
Label propagation	0.78	0.75	0.66	0.42
<b>EM-style approach</b>	<b>0.78</b>	<b>0.76</b>	<b>0.72</b>	<b>0.64</b>

# EM-Style approach outperforms all baselines

- Our Model: EM-Style Approach

- Yoga:
  - Accuracy: 78%
  - Macro-avg F1 score: 76%
- Keto:
  - Accuracy: 72%
  - Macro-avg F1 score: 64%

Model	Yoga		Keto	
	Accuracy	Macro-avg F1	Accuracy	Macro-avg F1
LSTM_Glove	0.51	0.45	0.72	0.43
Fine-tuned BERT	0.47	0.47	0.72	0.42
Label propagation	0.78	0.75	0.66	0.42
<b>EM-style approach</b>	<b>0.78</b>	<b>0.76</b>	<b>0.72</b>	<b>0.64</b>

- Baseline Models:

- Weakly Supervised Baseline:
  - Label Propagation
- Supervised Baseline:
  - LSTM\_Glove
  - Fine-tuned BERT

# Does Multiview Information Help?

Model	Yoga		Keto	
	Accuracy	Macro-avg F1	Accuracy	Macro-avg F1
Label propagation (des)	0.721	0.711	0.715	0.398
EM-style approach (des)	0.781	0.761	0.664	0.635
Label propagation (net)	0.573	0.572	0.644	0.384
EM-style approach (net)	0.670	0.657	0.707	0.617
Label propagation (des + net)	0.781	0.753	0.663	0.418
<b>EM-style approach (des + net)</b>	<b>0.782</b>	<b>0.763</b>	<b>0.722</b>	<b>0.642</b>

---

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

---

# Does Multiview Information Help?





Model	Yoga		Keto	
	Accuracy	Macro-avg F1	Accuracy	Macro-avg F1
Label propagation (des)	0.721	0.711	0.715	0.398
EM-style approach (des)	0.781	0.761	0.664	0.635
Label propagation (net)	0.573	0.572	0.644	0.384
EM-style approach (net)	0.670	0.657	0.707	0.617
Label propagation (des + net)	0.781	0.753	0.663	0.418
<b>EM-style approach (des + net)</b>	<b>0.782</b>	<b>0.763</b>	<b>0.722</b>	<b>0.642</b>

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

Multiview information improves prediction performance compared to the models using only either profile description or user network information.



# Roadmap

	Brief Introduction to Graph Embedding Model
	Dataset Collection and Annotation
	Automatic User Characterization
	User Type Analysis

# Tweets and Labels



(a) yoga: practitioner



(b) yoga: promotional

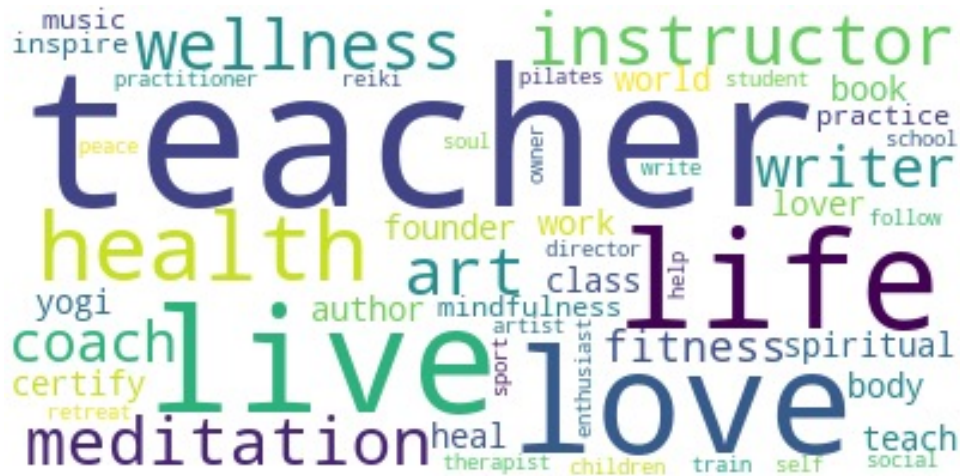


(c) keto: practitioner



(d) keto: promotional

## Profile Description and Labels



(a) yoga: practitioner



(b) yoga: promotional



(c) keto: practitioner



(d) keto: promotional



# Users' Sentiment Analysis



(a) yoga: practitioner



(b) keto: practitioner

# Summary of Contributions

- Formulate a novel problem of exploiting weak supervision for characterizing users in social media.
- Suggest a graph embedding based EM-style approach for learning and reasoning to construct like-minded users incrementally.
- Generate weak labels from user's profile description along with quantitative quality assessment.
- Conduct extensive experiments on real-world datasets to demonstrate the effectiveness of the model.

# THANK YOU 😊

**Slide:** [https://tunazislam.github.io/files/ICWSM22\\_yoga\\_keto.pdf](https://tunazislam.github.io/files/ICWSM22_yoga_keto.pdf)

## Questions?

**Tunazzina Islam**

Department of Computer Science,  
Purdue University, West Lafayette, IN.

Email: [islam32@purdue.edu](mailto:islam32@purdue.edu)

 <https://tunazislam.github.io/>

 [@Tunaz\\_Islam](https://twitter.com/Tunaz_Islam)

