# Analysis of Twitter Users' Lifestyle Choices using Joint Embedding Model

**Tunazzina Islam**, Dan Goldwasser
Department of Computer Science, Purdue University, West Lafayette, IN-47907

PURDUE UNIVERSITY

## Background

### Motivation

- Multiview representation of social media data.
- Construct coherent and contextualized user representations.
- Analyze users' activity type and motivation for specific lifestyle.

### Contribution

- Annotated dataset related to 'yoga' and 'keto diet'.
- A joint embedding model incorporating users' social and textual information.
- Extensive empirical experiments.
- Qualitative analysis to describe relationship between output labels and several different indicators, i.e., tweets, descriptions, location.

### Downstream Tasks

- Finding user type, i.e., Practitioner, Promotional, Other.
- Finding user motivation, i.e., Health, Spiritual, Other.

## Data

### Collection

- **Yoga data**
  - ~ **0.4 million** yoga-related tweets from Twitter.
  - **1298** users: at least 5 yoga-related tweets.
- **Keto data**
  - ~ **75k** keto-related tweets from Twitter .
  - **1300** users: at least 2 keto-related tweets.

### Pre-processing

- Convert to lower case.
- Remove URLs, smiley, emoji.
- Tokenize text using BERT and RoBERTa's wordpiece tokenizer.

### Annotation

- 1 annotator, with annotation instruction and examples provided.
- To calculate % agreement, 2 graduate students annotate a subset of tweets having inter-annotator agreement **64.7%**.

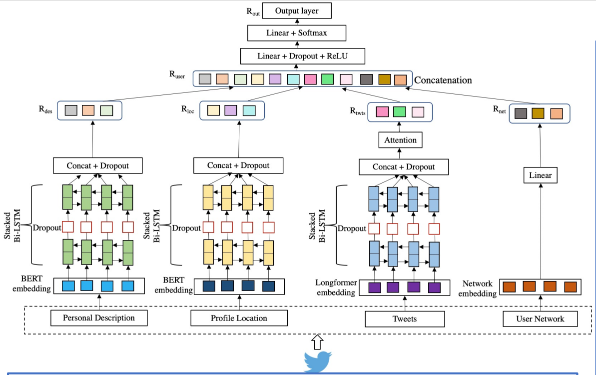
## Model

### Tweets Representation, $R_{twts}$

- Concatenate all yoga/keto-related tweets representing long documents.
- Embedding with longformer-base-4096.
- Pass to stacked Bi-LSTM with dropout (0.5).
- Hidden representation by concatenating forward and backward directions with dropout (0.5).
- Use context-aware attention.

### Metadata Representation, $R_{des}$ and $R_{loc}$

- Embedding with uncased BERTbase.
- Pass to stacked Bi-LSTM with dropout (0.5).
- Hidden representation by concatenating forward and backward directions with dropout (0.5).




### Network Representation, $R_{net}$

- Dense user network considering users from dataset if they are @-mentioned.
- Nodes: all users in the dataset.
- An edge: if either user mentions/retweets the other (from our data).
- Use Node2Vec for computing node embedding.
- Pass through a linear layer.

### User Representation, $R_{user}$

- Concatenate four representations: $R_{user} = [\, R_{des}, R_{loc}, R_{twts}, R_{net} \,]$
- Feed to a fully connected two-layer classifier activated by *ReLU* and *softmax*.
- Use dropout between individual neural network layers.
- SGD over shuffled mini-batches with Adam.
- Cross-entropy loss as the objective function for classification.

## Results

| Model | User type | | User motivation | |
|---|---|---|---|---|
| | Accuracy | Macro-avg F1 | Accuracy | Macro-avg F1 |
| Description | 0.694 | 0.611 | 0.707 | 0.523 |
| Location | 0.639 | 0.520 | 0.694 | 0.517 |
| Tweets | 0.795 | 0.704 | 0.786 | 0.595 |
| Network | 0.726 | 0.561 | 0.798 | 0.590 |
| Des_BF | 0.718 | 0.681 | 0.771 | 0.528 |
| Loc_BF | 0.679 | 0.606 | 0.695 | 0.476 |
| Twts_BF | 0.760 | 0.669 | 0.805 | 0.551 |
| Des + Loc | 0.734 | 0.653 | 0.806 | 0.661 |
| Des + Net | 0.808 | 0.702 | 0.823 | 0.653 |
| Des + Loc + Twt | 0.778 | 0.705 | 0.808 | 0.603 |
| Des + Loc + Net | 0.774 | 0.725 | 0.806 | 0.663 |
| Word2Vec based joint embedding | 0.790 | 0.742 | 0.844 | 0.610 |
| **Our Model** | **0.802** | **0.757** | **0.853** | **0.708** |

- **Yoga user type:** Accuracy: **80.2%**, Macro-avg F1: **75.7%**
- **Yoga user motivation:** Accuracy: 85.3%, Macro-avg F1: **70.8%**
- **Keto user type:** Accuracy: **71.9%**, Macro-avg F1: **67.6%**

## Analysis

### Top Hashtags

- In yoga dataset, the popular hashtag **#namaste** ('bow me you' or 'I bow to you'), **#gfyh** ('Go 4 Yoga Health'), **#mantra** ('vehicle of the mind').
- Keto diet is related to low carb high-fat diet having common hashtag **#lchf**.

### Relationship between Tweets and Labels

- Yoga practitioners' wordcloud: ***practice, love, pose, class, meditation, mind, mantra, thank, gfyh, yogaeverywhere***.
- Yoga promotionals' wordcloud: ***class, studio, come, train, teacher, workshop, free, mat, offer***.
- Other users mostly retweet and share news of yoga/yogi rather than directly practicing or promoting yoga. They have noticeable words such as ***rt, reiki, sadhguru, isha, yogaday*** in wordcloud.
- Keto practitioners' wordcloud: ***diet, low carb, fat, carnivore, ketosis, start, go, try, love, fast, protein, meat, egg***.
- Keto promotionals' wordcloud: ***recipe, paleo, weight loss, delicious, meal prep, money, healthy, organic, yummy***.
- Other keto users: ***rt, ketodietapp, ketogenic diet, ketone, low carb, cook, health benefit***.

### Relationship between Descriptions and Labels

- Yoga practitioners' wordcloud: ***yoga, teacher, health, fitness, meditation, lover, coach, founder, author, writer, instructor, certify***.
- Yoga promotionals' wordcloud: ***yoga, fitness, wellness, community, event, offer, free, product, market, business, program, design***.
- Users who practice yoga for health benefits have similar wordcloud to yoga practitioners' descriptions.
- Spiritually motivated yoga user having words like ***yoga, spiritual, spirituality, devotee, wisdom, peace, seeker, yogi, meditator, Indian***.
- Keto practitioners' wordcloud: ***keto, love, life, food, family***.
- Keto promotionals' wordcloud: ***food, health, keto, meal, recipe, product, online, free***.

### Relationship between Location and Labels

- We observe that we have more practitioners and promotional yoga users from the USA than the rest of the world.
- We find South-Asian users mostly retweet about yoga.
- We notice more 'others' users than practitioners in India.
- Most of the yoga users from India are motivated spiritually.
- For keto, we notice that our data is skewed towards the USA.

### Error Analysis

- Profile description, tweets, and network field contribute mainly to the classification task.
- Some prediction errors arise when description fields are absent or misleading.
- User location has relatively low accuracy and macro-avg F1 score according to ablation study.
- As Longformer supports sequences of length up to 4096, we might lose some information from tweets if the size of concatenated tweets > 4096.
- We construct @–mentioned network directly from retweets/mentions in tweets, which is less expensive to collect than the following network.

## Conclusion & Future Work

- We propose a BERT based joint embedding model that explicitly learns contextualized user representations by leveraging users' social and textual information.
- We show that our model outperforms multiple baselines.
- Besides yoga, we demonstrate that our model can effectively predict user type on another lifestyle choice, e.g., 'keto diet' and our approach is a general framework that can be adapted to other corpora.
- In the future, we aim to investigate our work to a broader impact like community detection based on different lifestyle decisions using minimal supervision.