Analysis of Twitter Users' Lifestyle Choices using Joint Embedding Model

Tunazzina Islam, Dan Goldwasser

Department of Computer Science

Purdue University, West Lafayette, IN

ICWSM 2021

Date: June 7-10, 2021

Motivation

- Multiview representation learning for constructing coherent and contextualized users' representations.
- A joint embedding model incorporating users' social and textual information to learn contextualized user representations for understanding their lifestyle choices.
- Use the model to analyze users' activity type and motivation.
- Tweets – (i) yoga, (ii) keto diet.

Objective

Challenge: Construct contextualized user representation relevant for characterizing nuanced as well as activity and lifestyle specific properties.

Advantages: General framework adapted to other corpora.
i.e., effectively predict user type on another lifestyle choice, e.g., 'keto diet'.

Contribution:

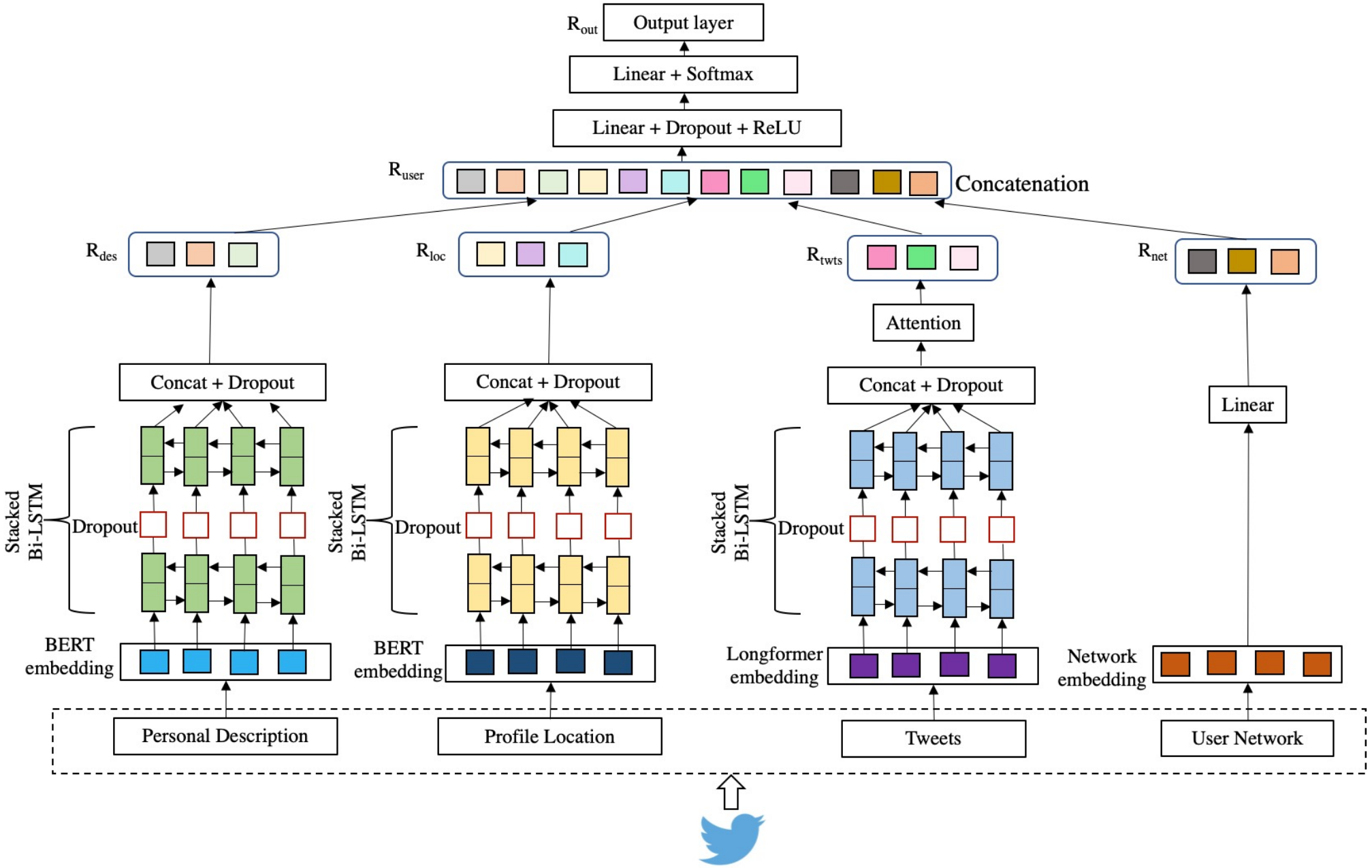
- Annotated dataset related to 'yoga' and 'keto diet'.
- Building a model for aggregating users' tweets as well as metadata and contextualizing this textual content with social information.
- Extensive empirical experiments and our model outperforms the baselines.
- Qualitative analysis aimed at describing the relationship between the output labels and several different indicators, including the tweets, profile descriptions, and location information.

Downstream Tasks

We demonstrate our model on two downstream tasks:

1) Finding user type
    1) Practitioner
    2) Promotional
    3) Other
2) Finding user motivation
    1) Health
    2) Spiritual
    3) Other

Model Architecture



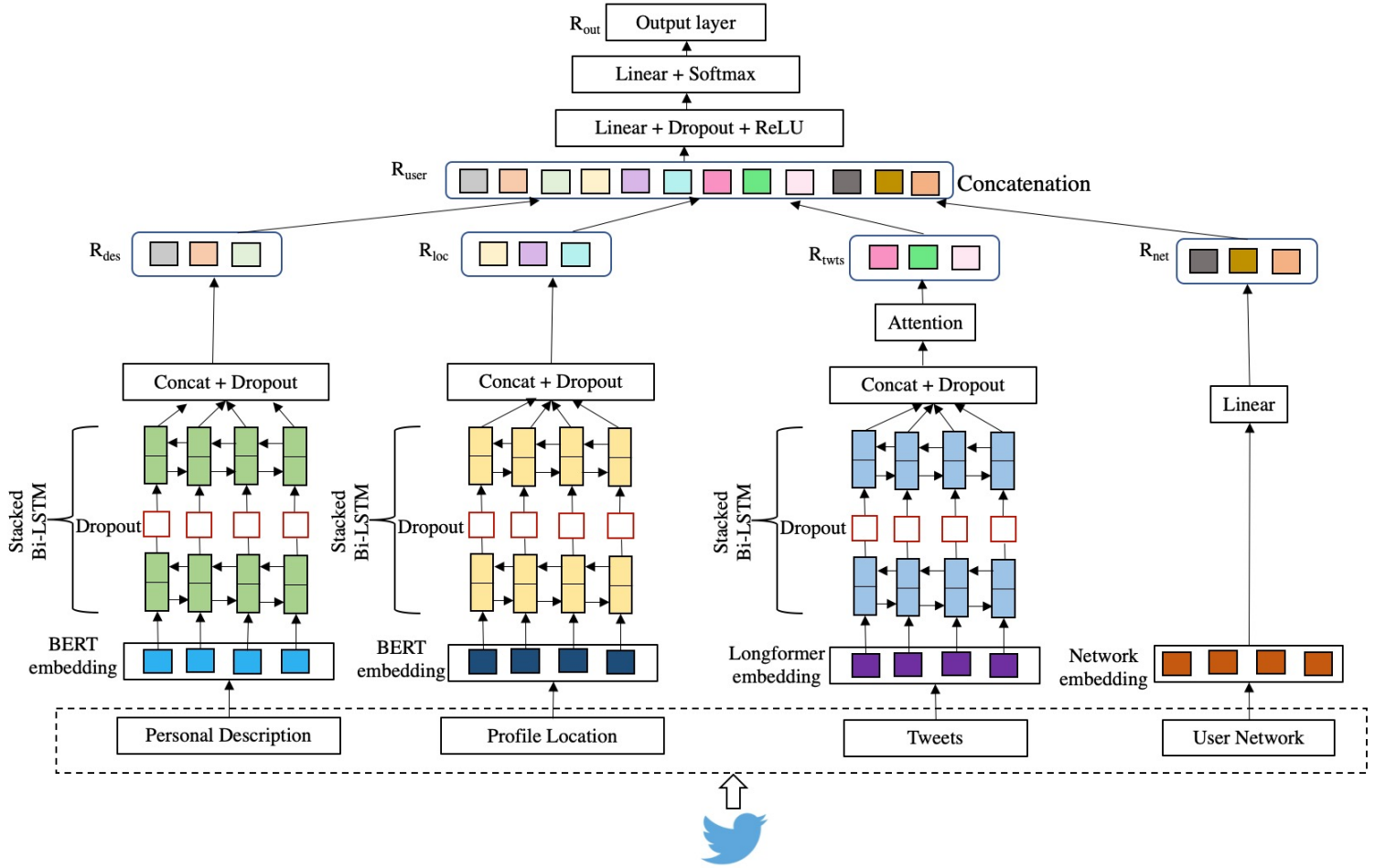
Model

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).
- Get the hidden representation of tweets by concatenating the 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).
- Get the hidden representation of metadata by concatenating the forward and backward directions with dropout (0.5).



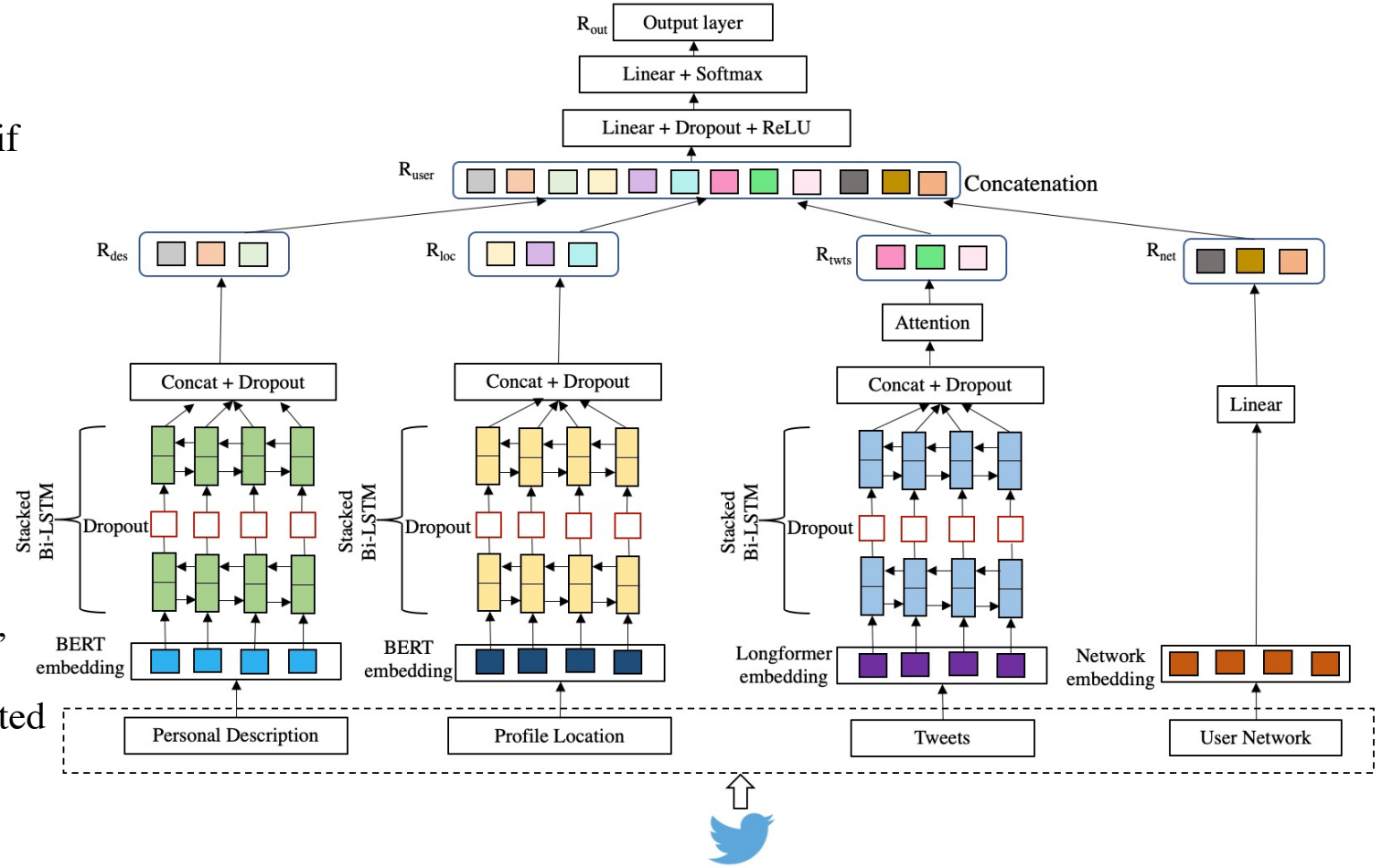
Model

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 optimizer.
- Cross-entropy loss as the objective function for classification.



Dataset

- Yoga data
  - ~ **0.4 million** yoga-related tweets from Twitter using Twitter streaming API (May to November of 2019) containing specific keywords.
  - **1298** users have at least **5** yoga-related tweet in their timelines.
  - ~ **3 million** of timeline tweets.
- Keto data
  - ~ **75k** keto-related tweets from Twitter (May to November of 2019) containing specific keywords.
  - **1300** users have at least **2** keto-related tweet in their timelines.
  - ~ **3.2 million** of timeline tweets.
- Pre-processing:
  - Convert to lower case.
  - Remove URLs, smiley, emoji.
  - Tokenize the 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% (substantial agreement).

Baseline Models

- User type and motivation detection baseline – 12 baselines
    1. Description only;
    2. Location only;
    3. Tweets only;
    4. Network only;
    5. BERT finetuned with Description (Des_BF);
    6. BERT fine-tuned with Location (Loc_BF);
    7. BERT fine-tuned with Tweets (Twts_BF);
    8. Joint embedding on description and location (Des + Loc);
    9. Joint embedding on description and network (Des + Net);
    10. Joint embedding on description, location, and tweets (Des + Loc + Twt);
    11. Joint embedding on description, location, and network (Des + Loc + Net);
    12. Word2Vec based joint embedding on description, location, tweets, and network.

Results

Our model outperforms the baselines.

Yoga:

- Accuracy (user type): 80.2%
- Macro-avg F1 score (user type): 75.7%
- Accuracy (user motivation): 85.3%
- Macro-avg F1 score (user motivation): 70.8%

Keto:

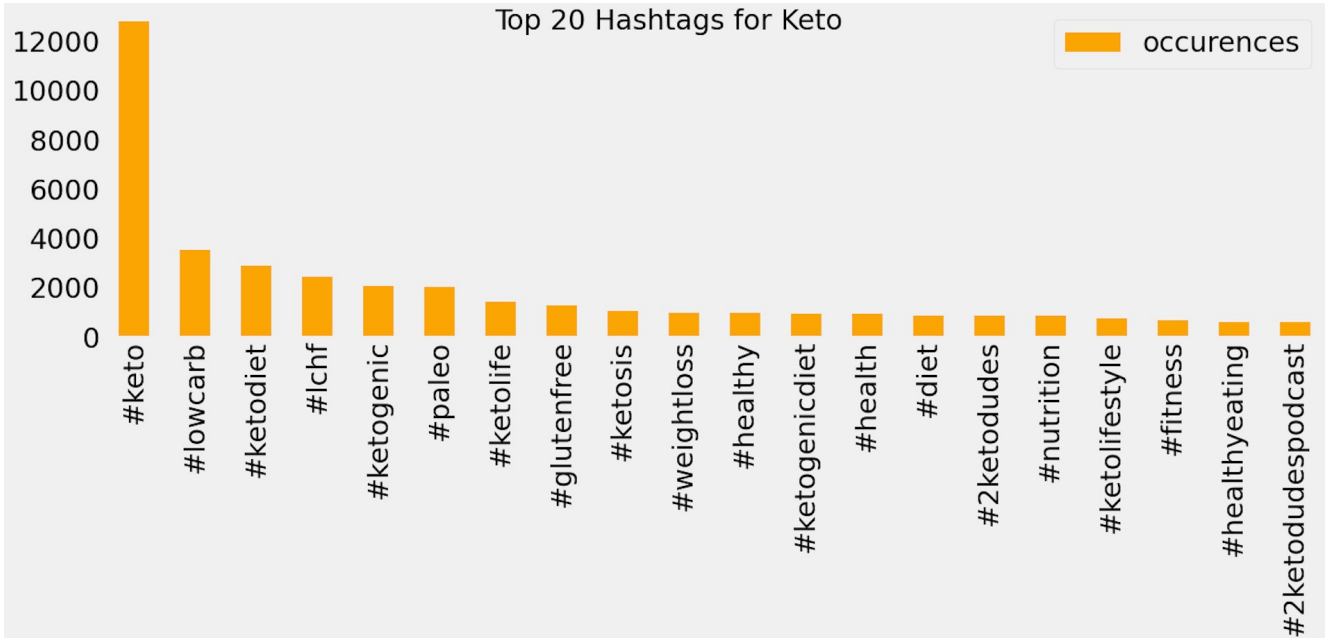
- Accuracy (user type): 71.9%
- Macro-avg F1 score (user type): 67.6%

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**

Table 2: Performance comparisons on yoga data.

Analysis

- Top Hashtags



Analysis

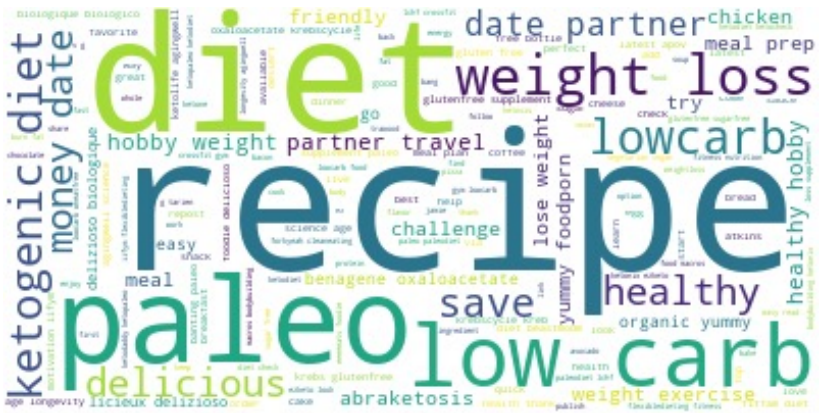
- Relationship between Tweets and Labels:

(a) yoga: practitioner

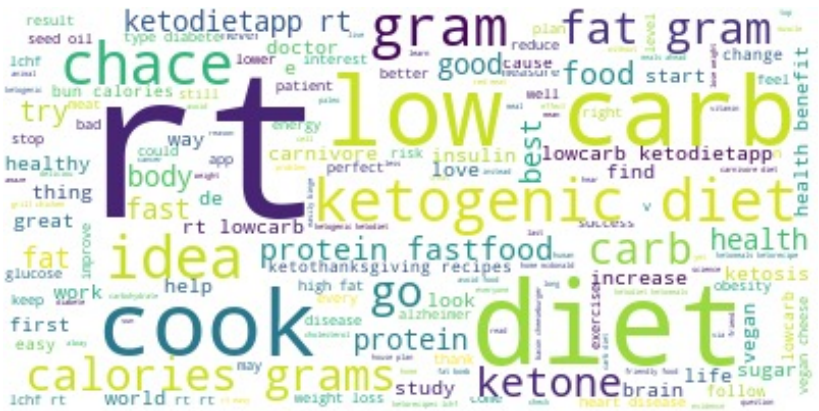
(b) yoga: promotional

(c) yoga: other

(d) keto: practitioner



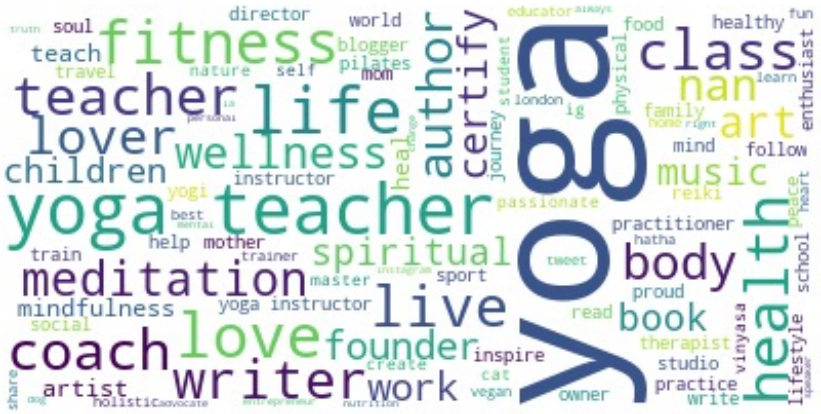
(e) keto: promotional



(f) keto: other

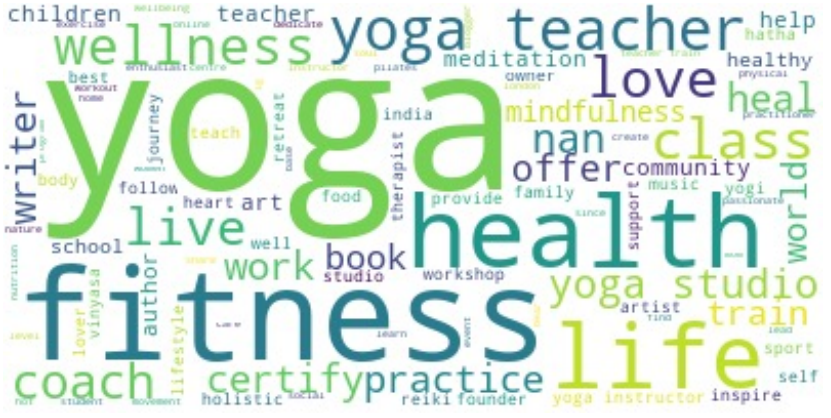
Analysis

- Relationship between Descriptions and Labels:

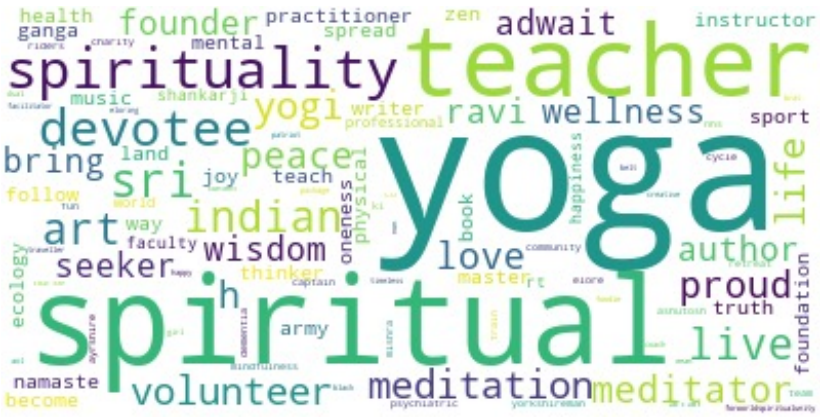


(a) yoga: practitioner

(b) yoga: promotional



(c) yoga: health motivation



(d) yoga: spiritual motivation

(e) keto: practitioner

(f) keto: promotional

Analysis

- Relationship between Location and Labels
  - We observe 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

- 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, some information from tweets might be missing if the size of concatenated tweets > 4096.
- Constructing @−mentioned network directly from retweets/mentions in tweets instead of collecting the following network (expensive).

Conclusion and Future Work

- BERT based joint embedding model that explicitly learns contextualized user representations by leveraging users' social and textual information.
- Our model outperforms multiple baselines.
- Our model can effectively predict user type on another lifestyle choice, e.g., 'keto diet'.
- 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.

THANK YOU ☺

Slide: https://tunazislam.github.io/files/ICWSM21_yoga_keto.pdf

Questions?

Tunazzina Islam

Department of Computer Science,
Purdue University, West Lafayette, IN.
Email: **islam32@purdue.edu**

https://tunazislam.github.io/

@Tunaz_Islam

PURDUE UNIVERSITY