From Tweets to Wellness: Wellness Event Detection from Twitter Streams

Mohammad Akbari [a,b], Xia Hu[c], Nie Liqiang[a], Tat-Seng Chua [a,b]
[a] NUS Graduate School for Integrative Sciences and Engineering, National University of Singapore, Singapore
[b] School of Computing, , National University of Singapore, Singapore
[c] Department of Computer Science & Engineering, Texas A&M University, USA
akbari@u.nus.edu, hu@cse.tamu.edu, nieliqiang@gmail.com, chuats@comp.nus.edu.sg

Abstract

Social media platforms have become the most popular means for users to share what is happening around them. The abundance and growing usage of social media has resulted in a large repository of users' social posts, which provides a stethoscope for inferring individuals' lifestyle and wellness. As users' social accounts implicitly reflect their habits, preferences, and feelings, it is feasible for us to monitor and understand the wellness of users by harvesting social media data towards a healthier lifestyle. As a first step towards accomplishing this goal, we propose to automatically extract wellness events from users' published social contents. Existing approaches for event extraction are not applicable to personal wellness events due to its domain nature characterized by plenty of noise and variety in data, insufficient samples, and inter-relation among events. To tackle these problems, we propose an optimization learning framework that utilizes the content information of microblogging messages as well as the relations between event categories. By imposing a sparse constraint on the learning model, we also tackle the problems arising from noise and variation in microblogging texts. Experimental results on a real-world dataset from Twitter have demonstrated the superior performance of our framework.

Recent years have witnessed the revolutionary changes brought by the development of social media services through which individuals extensively share information, express ideas, and construct social communities. These changes can advance many disciplines and industries, and health is no exception (Nie et al. 2015; Lee et al. 2014). In such a context, many users are keen to share their wellness information on social platforms such as Twitter and Facebook (Hawn 2009; Yang et al. 2014; Dos Reis and Culotta 2015; Paul et al. 2015). Take diabetes as an example; diabetic patients not only share about events happening around them but also frequently post about their current health conditions, medication, and the outcomes of medications. For instance, they frequently post the latest values of their blood glucose, diet, and exercises using "*#diabetes*" and "*#BGnow*" hashtags on Twitter. This provides new opportunities to understand individuals' wellness that can be used to assist them in managing their health in a scope that previously was impossible. As a first step towards accomplishing this end, we propose to automatically extract wellness events from users' published social contents.

Copyright © 2016, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Extraction of personal wellness events (PWEs) will provide significant insights about individual's wellness and community lifestyle behaviours. At the individual level, it can summarize the past wellness events of individuals which significantly facilitate lifestyle management through coarse and fine-grained browsing. PWE summary can be useful for downstream applications such as user health profiling, personalized lifestyle assistant, and targeted online advertising. Take diet as an example; if one diabetic person consumes a lot of carbohydrates, the system can offer diet suggestion. At the community level, accumulating the wellness information of a large set of individuals makes it feasible to analyze and understand the lifestyle patterns and wellness of social groups in a scale that was impossible with traditional methods in terms of both time and cost.

Despite its value and significance, extracting PWEs from social media services has not been fully investigated due to the following challenges. First, the language used in social media is highly varied, informal, and full of slang words. Second, PWEs are relatively rare in social media posts as users tend to post their personal significant events together with lots of trivialities and other public events (Li et al. 2014). As a result, wellness events are buried among other contents produced by the users and their social connections. Identifying wellness events from a huge volume of other non-wellness events poses a big challenge. As a result, even a large annotated dataset might contain just a few examples of PWE categories. Third, the structure of wellness events exhibits a hierarchical taxonomy as shown in Table 1. Indeed, events under the same category are closely related. For instance, clinical tests are much more related to treatment, than running. These events may share some features such as entities, attributes and relations, which makes the problem arduous. How to mathematically model such relations and integrate them into a learning framework remains a challenge.

In health sciences, it has been intensively studied and well-established that physical activities, diet planning and taking prescribed medications are the key therapeutic

Table 1: Taxonomy of wellness events with exemplar tweets.

Event	Sub Event	Example
Diet	Meals	Dinner just salad
	Alcoholic Beverages	Too much drink in party
	Non-alcoholic Beverages	Talking about hot chocolates, I might just go and make myself one :D
	Snacks	found Taylor's pretzels in my backpack and I'm so happy wow
	Fruit	almost eat all the strawberries
	Others	Eat 20g carbs and go fo running
Exercise	Walking	20 mins walk around office..
	Running	after 1 hour run #bgnow 130
	Biking	I just finished 1 hour biking
	Swimming	BGnow 95, thanks swimming pool
	Others	Shopping and having a little dinner URL
Health	Examinations	#BGnow 100
	Symptoms	Feel too much Fatigue
	Treatment	ate great oatmeal, toast, and eggs. Had 1 unit

treatments of many diseases (Pastors et al. 2002; Hu 2011). Further, unhealthy lifestyle behaviours such as unhealthy dietary habits, sedentary lifestyle, and the harmful consumption of alcohol are mainly related to the risk factors of noncommunicable diseases (NCDs) ranked as the leading cause of disability-adjusted life years (DALYs) (Lim et al. 2013; Association and others 2014). Therefore, the primary aim of the General Assembly of the United Nations on NCDs in 2011 was to reduce the level of exposure of individuals and population to NCDs' risk factors and strengthen the capacity of individuals to follow lifestyle patterns that foster good health[1].

As a first step towards accomplishing this end, we propose a supervised model to extract PWEs from social media posts of a given user and categorize them into a taxonomy as shown in Table 1. In particular, we propose an optimization learning framework that utilizes the content information of microblogging messages as well as the relations among event categories. We seamlessly incorporate these two types of information into a sparse learning framework to tackle noisy texts in microblogs.

The main contributions of this paper are threefold:

- As far as we know, this is the first work on personal wellness event extraction from social media posts of individuals. Although experiments were performed on diabetic users who use Twitter microblogging platform, it is easily extendable to other diseases.
- We present a novel supervised model for wellness event extraction that takes task relatedness into account to learn task-specific and task-shared features.
- We construct a large-scale diabetes dataset by automatically extracting lifestyle and wellness related short messages and manually constructing the ground-truth labels. We plan to release this dataset to facilitate others in reproducing our experiments as well as verifying their ideas[2].

Problem Statement

The problem we study in this paper is different from traditional event detection since the latter normally focuses on detecting and constructing an evolutionary timeline of public events (Becker, Naaman, and Gravano 2011; Meladianos et al. 2015). Moreover, they assume that events are independent and hence only consider content information to identify event categories. In this section, we first present the notations and then formally define the problem of PWE detection from individuals' social media accounts.

We use boldface uppercase letters (e.g., $\mathbf{A}$) to denote matrices, boldface lowercase letters (e.g., $\mathbf{a}$) to denote vectors, and lowercase letters (e.g., a) to denote scalars. The entry at the i-th row and j-th column of a matrix $\mathbf{A}$ is denoted as $\mathbf{A}_{ij}$. $\mathbf{A}_{i*}$ and $\mathbf{A}_{*j}$ denote the i-th row and j-th column of a matrix $\mathbf{A}$, respectively. $\|\mathbf{A}\|_1$ is the ℓ_1-norm and $\|\mathbf{A}\|_F^2$ is the Frobenius norm of matrix $\mathbf{A}$.

Suppose that there are M wellness events and let $\mathcal{C} = \{c_1, c_2, ..., c_M\}$ be the set of class labels. Given a corpus $\mathcal{P} = \{p_1, p_2, ..., p_N\}$ composed of N different training samples. Each training sample $p_i = (\mathbf{x}_i, \mathbf{y}_i)$ consists of a message content vector denoted by $\mathbf{x}_i \in \mathbb{R}^J$ and the corresponding event label vector denoted by $\mathbf{y}_i \in \mathbb{R}^M$. Let $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N]^T \in \mathbb{R}^{N\times J}$ be the matrix representing training data and $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_N]^T \in \mathbb{R}^{N\times M}$ be the matrix of labels. Our learning task is to find a mapping function from feature space $\mathbf{X}$ to label $\mathbf{Y}$.

With the notation above, we formally define the personal wellness event detection problem as: *Given a sequence of microblog messages $\mathcal{P}$ with their content $\mathbf{X}$, and the corresponding event labels $\mathbf{Y}$, we aim to learn a model $\mathbf{W}$ to automatically assign events' labels for unseen messages (i.e., test data).*

Wellness Event Categorization

In essence, two characteristics of personal wellness event detection are: 1) training data is sparse; and 2) event categories are deeply inter-related. Their associated challenges are: a) which events are related in problem domain; and b) how to incorporate event relations into the learning framework to infer a more effective learning model. In this section, we first explain how to formulate the problem of PWE detection as a multi-task learning (MTL) framework which utilizes the content information of microblogging texts as well as captures the relation between the event categories into an integrated learning framework. We samelessly integrate these two types of information into a state-of-the-art framework and turn the integrated framework into an optimization problem. We then demonstrate how to find the solution of the problem with an efficient framework.

Modeling Content Information

Traditionally, supervised learning is widely used to infer topics of text documents. A straight forward way for event detection is to learn a supervised model based on labeled data, and apply the model to detect the topics of each microblogging post. However, compared with textual documents in traditional media, a distinct feature of texts in microblogging platforms is that they are noisy and short (Chen et al. 2013; Hu, Tang, and Liu 2014), which give rise

[1] http://www.un.org/en/ga/ncdmeeting2011/

[2] Available at www.comp.nus.edu.sg/~a0103416.

to two issues. First, text representation models, like "Bag of Words" (BoW) and n-grams, lead to a high-dimension feature space due to the variety of words. Second, the posts are too short and noisy making the representation very sparse. To mitigate these problems, we propose a sparse model to perform classification of feature space.

Assume that we have M wellness events, and view each event as one task. Formally, we have M tasks $\{T_1, T_2, ..., T_M\}$ in the given training set $\mathcal{P}$. The prediction for each task t is given by $f_t(\mathbf{x}; \mathbf{w}_t) = \mathbf{x}^T \mathbf{w}_t$, where $\mathbf{w}_t$ is the coefficient for task t. The weight matrix of all M tasks can be denoted as $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_M] \in \mathbb{R}^{J \times M}$. Matrix $\mathbf{W}$ can be inferred from the training data by solving the following optimization problem:

$$\arg\min_{\mathbf{W}} \mathcal{L}(\mathbf{X}, \mathbf{W}, \mathbf{Y}) + \Phi(\mathbf{W}), \tag{1}$$

where $\mathcal{L}(.)$ is the loss function, and $\Phi(\mathbf{W})$ is a regularizer which controls the complexity of the model to prevent overfitting and selects discriminant features. This formulation is a sparse supervised method, where the data instances are independent and identically distributed (*i.i.d*), and the tasks are independent.

The loss function $\mathcal{L}(\mathbf{X}, \mathbf{W}, \mathbf{Y})$ is defined as logistic loss in this work,

$$\sum_{t=1}^{M} \sum_{i=1}^{N} log(1 + exp(-y_i^t f_t(\mathbf{x_i}, \mathbf{w_t}))), \tag{2}$$

where $y_i^t \in \{-1, 1\}$ is the true label indicating the relevance of i-th sample to the t-th task. Note that each sample can fall into multiple categories. For instance, "banana bread in the oven, mmmmm! lets just enjoy this #bgnow 70!" is related to meals and health examination categories at the same time.

To select discriminant features and control the complexity of our model, we define $\Phi(\mathbf{W})$ as follows,

$$\Phi(\mathbf{W}) = \alpha \|\mathbf{W}\|_F^2 + \beta \|\mathbf{W}\|_1, \tag{3}$$

where, α and β are positive regularizer parameters. In the defined regularizer $\Phi(\mathbf{W})$, the first term, i.e. Frobenius-norm, controls the generalization performance of the model and the second term, i.e. ℓ_1-norm, leads to a sparse representation for the texts, which performs feature selection to reduce the effects of noisy features. Thus $\Phi(\mathbf{W})$ performs a kind of continuous feature selection as well as controls the complexity of the model (Ruvolo and Eaton 2014; Song et al. 2015).

Modeling Events Relations

Recall that PWE detection has two characteristics: 1) some events are more related to each other while differ from others, and similar events might share some features. For example, "walking" shares some features with "running" since the context of two events are similar. However, it greatly differs from "meals"; 2) the dimension of feature space is usually very high. In fact, some features are not discriminative enough for wellness event detection. This motivates us to propose a graph-guided multi-task learning model, which is capable of capturing the relatedness among tasks to learn task-shared features as well as the task-specific features. The hope is that common information relevant to prediction can be shared among tasks and jointly learning of tasks' models leads to a better generalization performance as compared to independently learning each task. A major challenge hence is how to control the sharing of information among tasks so that it leads to close models for related tasks while unrelated tasks do not end up influencing each other. Therefore, the key assumption for our model is that tasks are assumed to be related to each other with different weights and the parameters of two related tasks are close to each other in ℓ_2 norm sense.

Based on the above discussion, to incorporate task relations into event detection, we assume that the task relationships can be represented using a graph structure G, where each node represents one task and each edge connects two related tasks. The weight of each edge $r(t_i, t_j)$ reflects the relation strength between task i and j. Given a graph G, we can formulate the task relations as minimizing the following objective function $\Omega(\mathbf{W})$,

$$\begin{aligned}\Omega(\mathbf{W}) &= \lambda \sum_{t_i, t_j \in \mathcal{E}} r(t_i, t_j) \|\mathbf{W}_{*i} - \mathbf{W}_{*j}\|_2^2 \\ &= \lambda \operatorname{Tr}(\mathbf{W}(\mathbf{V} - \mathbf{R})\mathbf{W}^T) = \lambda \operatorname{Tr}(\mathbf{W} \boldsymbol{\Delta} \mathbf{W}^T),\end{aligned} \tag{4}$$

where $\mathcal{E}$ contains all the edges of graph G, and $\boldsymbol{\Delta} = \mathbf{V} - \mathbf{R}$ is the graph Laplacian matrix (Nie et al. 2014a; Akbari, Nie, and Chua 2015), where $\mathbf{R} \in \mathbb{R}^{M \times M}$ is the task relatedness matrix. $\mathbf{R}_{ij} = r(t_i, t_j)$ indicates the relation strength between task i, and j and $\mathbf{R}_{ij} = 0$, otherwise. $\mathbf{V} = diag(\mathbf{V}_{jj})$ is a diagonal matrix with $\mathbf{V}_{jj} = \sum_{i=1}^{M} r(t_i, t_j)$. The regularizer parameter λ controls the impact of relations amongst tasks in learning process.

To construct the graph, we utilize a fully automated approach based on the model learnt from the relaxed multi-task problem. Following the idea discussed in (Kim and Xing 2009), we first train a MTL model with Lasso regularizer to compute the model for each tasks t_i and then compute the pairwise correlation between distinct tasks. We simply create an edge between each pair of tasks which have correlation above a defined threshold ρ. We set the threshold to $\rho = 0.7$ since it leads to the best performance in our experiments.

The optimization framework, which integrates content information and event relation information into the learning process, is defined by the integration of Eq. (1), through Eq. (4) as following objective function, $O(\mathbf{W})$,

$$\arg\min_{\mathbf{W}} \mathcal{L}(\mathbf{X}, \mathbf{W}, \mathbf{Y}) + \Phi(\mathbf{W}) + \Omega(\mathbf{W}), \tag{5}$$

where the first and second terms are to consider content information and perform regularization to avoid overfitting, respectively. The third term, i.e. $\Omega(.)$, captures tasks relatedness to learn task-shared features.

Optimization

The objective function $O(\mathbf{W})$ (i.e., Eq. (5)) is non-smooth since it is the composition of a smooth term and a non-smooth term, i.e. ℓ_1 penalty, and gradient descent method is not available to solve the formulation. In this section, we introduce an efficient algorithm to solve the optimization problem.

Inspired by (Nesterov 2004; Chen et al. 2009), we propose to solve the non-smooth optimization problem in Eq. (5) by optimizing its equivalent smooth convex reformulation. We hence derive an smooth reformulation of Eq. (5) by moving the non-smooth part, i.e. ℓ_1 norm, to the constraint.

Theorem 1. *Let $\mathcal{L}(\mathbf{X}, \mathbf{W}, \mathbf{Y})$ be a smooth convex loss function. Then Eq. (5) can be reformulated as the following ℓ_1-ball constrained smooth convex optimization problem:*

$$\arg\min_{\mathbf{W}\in\mathbf{Z}} f(\mathbf{W}) = \mathcal{L}(\mathbf{X}, \mathbf{W}, \mathbf{Y}) + \lambda\,\mathrm{Tr}(\mathbf{W}\boldsymbol{\Delta}\mathbf{W}^T) + \alpha\|\mathbf{W}\|_F^2, \tag{6}$$

where,

$$\mathbf{Z} = \{\mathbf{W} \,|\, \|\mathbf{W}\|_1 \leq z\}, \tag{7}$$

$z \geq 0$ is the radius of the ℓ_1-ball and there was a one-to-one correspondence between β and z [3].

We now find the solution for Eq. (6), which is equivalent to our optimization problem in Eq. (5). To solve the problem, we first consider the optimization problem without the constraint on $\mathbf{Z}$ which is defined as:

$$\arg\min_{\mathbf{W}} f(\mathbf{W}). \tag{8}$$

The solution to this problem can be computed from the gradient descent method which iteratively updates $\mathbf{W}_{i+1}$ in each step as follows,

$$\mathbf{W}_{i+1} = \mathbf{W}_i - \frac{1}{\gamma_i}\nabla f(\mathbf{W}_i), \tag{9}$$

where γ_i is the step size and it is determined by line search according to Armijo-Goldstein rule. The smooth part of the optimization problem can be reformulated equivalently as a proximal regularization of the linearized function $f(\mathbf{W})$ at $\mathbf{W}_i$ as,

$$W_{i+1} = \arg\min_{\mathbf{W}} \mathcal{M}_{\gamma_i,\mathbf{S}_i}(\mathbf{W}), \tag{10}$$

where,

$$\mathcal{M}_{\gamma,\mathbf{S}_i}(\mathbf{W}) = f(\mathbf{S}_i) + \langle\nabla f(\mathbf{S}_i), \mathbf{W} - \mathbf{S}_i\rangle + \frac{\gamma_i}{2}\|\mathbf{W} - \mathbf{S}_i\|_F^2, \tag{11}$$

where $\mathbf{S}_i$ is computed based on the past solutions by $\mathbf{S}_i = \mathbf{W}_i + \tau_i(\mathbf{W}_i - \mathbf{W}_{i-1})$. Eq. (11) can be rewritten as,

$$\arg\min_{\mathbf{W}} \mathcal{M}_{\gamma_i,\mathbf{S}_i}(\mathbf{W}) = \arg\min_{\mathbf{W}} \left(\frac{1}{2}\|\mathbf{W} - (\mathbf{S}_i - \frac{1}{\gamma_i}\nabla f(\mathbf{S}_i))\|_F^2\right) \tag{12}$$

By ignoring terms that are independent of $\mathbf{W}$ the objective function boils down to:

$$\mathbf{W}_{i+1} = \arg\min_{\mathbf{W}} \|\mathbf{W} - \mathbf{U}_i\|_F^2, \tag{13}$$

where $\mathbf{U}_i = \mathbf{S}_i - \frac{1}{\gamma_i}\nabla f(\mathbf{S}_i)$ and indeed the solution of $\mathbf{W}$ is the Euclidian projection of $\mathbf{U}_i$ on $\mathbf{Z}$. The overall optimization process can be described in Algorithm 1.

[3] The proof of the Theorem is available at www.comp.nus.edu.sg/~a0103416

Algorithm 1: Optimization algorithm of Eq. (5)

Input: $\mathbf{W}_0$, $\gamma_0 \in \mathbb{R}$, and $q =$ max iteration.
Output: $\mathbf{W}$.
Set $\mathbf{W}_1 = \mathbf{W}_0, t_{-1} = 0$, and $t_0 = 1$.
for $i = 1$ *to* q **do**
  Set $\tau_i = (t_{i-2} - 1)/t_{i-1}$.
  Set $\mathbf{S}_i = \mathbf{W}_i + \tau_i(\mathbf{W}_i - \mathbf{W}_{i-1})$.
  while *true* **do**
    Compute $\mathbf{W}^* = \arg\min_{\mathbf{W}} \mathcal{M}_{\gamma_i,\mathbf{S}_i}(\mathbf{W})$
    if $f(\mathbf{W}^*) \leq \mathcal{M}_{\gamma_i,\mathbf{S}_i}(\mathbf{W})$ **then**
      break
    else
      Set $\gamma_i = \gamma_i \times 2$
  Set $\mathbf{W}_{i+1} = \mathbf{W}^*$ and $\gamma_{i+1} = \gamma_i$.
  Set $t_i = \frac{1+\sqrt{1+4t_{i-1}^2}}{2}$.
Set $\mathbf{W} = \mathbf{W}_{i+1}$.

Experiments

In this section, we present the experimental details to verify the effectiveness of the proposed framework. We conduct experiments to answer the following questions that help to validate the framework:

1. How does the proposed framework perform as compared to other state-of-the-art baselines?
2. How well the selected features discriminate PWEs?
3. How sensitive is our model to the involved parameters?

Dataset Description

Recall that one of the main problems of this research is the lack of training data. According to our statistics, the wellness-oriented tweets are only less than 5% of all the messages posted by the chronic disease sufferers, and this value could be much smaller for healthy users. Therefore, we utilize a bootstrapping method to harvest the tweets corresponding to wellness events. We then manually label this tweet pool to construct our ground truth for training and evaluation.

Wellness event categories. Inspired by (Shelley 2005; Teodoro and Naaman 2013), we arrive at three high-level wellness categories, namely, diet, exercise & activities (exercise for brevity), and health as shown in Table 1. Under each high-level event category, we further organize specific sub-events which constructs a taxonomy comprises 14 distinct wellness events. We also define a null class for non-wellness events indicating a message is not directly related to any defined wellness event categories.

Assigning event labels. We observed that different wellness events place emphasis on different hashtags and words. For instance, we observed that "#dwalk" regularly appears in walking related posts. Inspired by (Mintz et al. 2009; Gupta and Manning 2014), we adopted a bootstrapping approach to select a set of tweet related to each wellness event. To do so, we first selected some representative seed words for each wellness events by verifying top frequent keywords of each category. We then gathered tweets explicitly involving these seed words.

Table 2: Statistics of the Dataset

	All samples	Positive samples
Posts on Diet	1979	710
Posts of Exercise	2771	1234
Posts on Health	8802	1300
Total Number of Posts	$13,552$	$3,244$

However, the collected tweets are weakly related to events and are full of noises. For instance, the tweet *"I love music,it has a voice for every walk of life,every emotion,every bit of love"* even containing the word "walk", but it is not a relevant one. To filter irrelevant tweets, we defined patterns in local context of each seed word. We applied the bootstrapping approach of (Thelen and Riloff 2002) to extend the set of keywords and patterns and collected more positive samples pertaining to wellness events. We stopped bootstrapping after 10 iterations since it often failed to find more positive candidates.

To construct the dataset, we first crawled a set of users who used #BGnow hashtag in their tweets. This hashtag is very popular among diabetic patients to post information about diabetes and their health states. In this way, we gathered $2,500$ different diabetes users. We removed accounts which had high daily traffic to avoid spammers. This filtering process resulted in $1,987$ diabetic users. We then crawled all historical tweets of these users using Twitter API, resulting in a set of about 3 million tweets. We applied the aforementioned bootstrapping procedure to find candidate tweets to construct dataset, which resulted in $13,552$ tweets. We labelled all the tweets based on the wellness events as shown in Table 1. For each given event, we regarded tweets labelled with its class as the positive training samples, and randomly selected negative samples from other events. Examples of the positive and negative tweets for the event "walking" are given below:

Positive 3 litres of water and 4 miles of walking I am feeling super refreshed...thank god!!

Negative Further evidence of the benefits of exercise for people with type 2 #diabetes URL #doc (Error: General Health Information)

Table 2 shows the statistics of our dataset. In total, our training set consists of approximately $3,000$ tweets corresponding to different wellness events. We also randomly selected about $3,000$ non-wellness tweets to be used as positive samples for the null class (non-wellness events). We intentionally selected more samples for null class due to the imbalance nature of events. We divided the dataset into two sets based on their posting times. In particular, tweets that were posted before May 2015 were utilized to train our model; while those posted from May to July 2015 were used for evaluation process.

We engaged another annotator to manually examine about $3,000$ messages. The inter-agreement between annotator was 0.857 with the *Cohen* κ metric, which verifies a substantial agreement between annotators.

Feature Settings

The following set of features were extracted to represent each tweet from the context and linguistic aspects:

- **NGrams**: We extracted unigrams and bigrams from Twitter messages since they are commonly used to represent user-generated contents (Hu and Liu 2012).
- **NE**: As shown in (Li et al. 2014; Ritter et al. 2012), the presence of named entities is a very useful indication of events in social media texts. We hence utilized named entities as another feature to represent social media messages (Ritter et al. 2012).
- **Gazetteer**: We utilized two set of gazetteers: food and drink gazetteers from (Abbar, Mejova, and Weber 2015) and time gazetteers from LIWC's time category (Pennebaker, Francis, and Booth 2001).
- **Modality**: Users frequently share general thoughts, wishes, and opinions in social platforms (Li et al. 2014). To filter these posts from those really reporting an event, we utilized modal verbs, like may and could, as an indicator of non-event information.

On Performance Evaluation

We conducted experiments to compare the performance of our model with other state-of-the-art approaches:

- **Alan**12: Event extraction method of (Ritter et al. 2012) which learns a latent model to uncover an appropriate event types based on available data.
- **SVM**: We trained a binary classifier for each event.
- **Lasso**: Logistic regression model with Lasso regularizer, i.e. ℓ_1 term (Tibshirani 1996).
- **GL-MTL**: Group Lasso regularizer with $\ell_{1/2}$ norm penalty for joint feature selection (Nie et al. 2010), which only encodes group sparsity.
- **TN-MTL**: Trace Norm Regularized MTL (Obozinski, Taskar, and Jordan 2010), which assumes that all tasks are related in a low dimensional subspace.
- **gMTL**: Our proposed wellness event detection model.

Table 3: Performance comparison among models.

Method	Precision	Recall	F-1 score
Alan12	62.70	48.10	54.44
SVM	83.05	79.65	81.31
Lasso	80.45	79.21	79.82
GL-MTL	84.37	80.72	82.50
TN-MTL	83.22	78.85	80.98
gMTL	87.15	82.69	84.86

For each method mentioned above, the respective parameters were carefully tuned based on 5-fold cross validation on the training set and the parameters with the best performance were used to report the final comparison results. The overall performance is shown in Table 3 in terms of precision, recall, and F-1 score metrics.

We can observe that all MTL methods outperform **Alan**12, **SVM** and **Lasso** in terms of precision with a substantial improvement over **Alan**12. The main reason is that event discovery methods mostly focus on detecting general events or major personal events (Zhou, Chen, and He 2015). These events are discussed bursty and highly

Table 4: Average performance of PWE detection on different feature setting.

Method	Precision	F-1 score
NGrams (Baseline)	82.70	81.06
Baseline+Gaz *	84.31	82.31
Baseline+Gaz+NE	86.85	84.04
All *	87.15	84.86

connected to specific name entities such as organizations, persons, and locations. However, PWEs are merely focus of individuals' local circles and may not be significantly related to any specific name entities. This hinders latent model to find representative latent topics from data. Among the multi-task approaches, **gMTL** achieves the best performance as compared to others. It verifies that there exists relationships among events and such relatedness can boost the learning performance. **GL-Lasso** achieves higher performance as compared to **Lasso** and **TN-MTL** since it tries to jointly learn features which resulted in better generalization. This verifies that sharing samples among distinct task alleviates the data scarcity problem as pointed out by previous studies (Ruvolo and Eaton 2014; Xu et al. 2015). The proposed **gMTL** model outperforms other methods by 2%-7% since it encodes the task relatedness and group sparsity. By sharing samples between different tasks, i.e. event categories, **gMTL** simultaneously learns task-shared and task-specific features as well as mitigates the problem of data scarcity.

On Feature Comparison

We also conducted an experiment to evaluate the effects of different features for PWE detection, as shown in Table 4. To conduct the study, we considered **NGram** feature as a baseline feature since it has been shown in many studies to have good performance (Tang et al. 2014; Nie et al. 2014b). We then added each distinct feature from the feature set and reported the average performance over all event categories. We also performed significant test to validate the importance of different features. We used the asterisk mark (*) to indicate significant improvement over the baseline.

As Table 4 shows, **NGram** and **Gaz** are important features for PWE detection. The reason might be that **NGram** represents the context information of messages and food gazetteer feature is a very effective indicator of events related to food and drink category which filter out many irrelevant samples. However, adding name entities, i.e. **NE**, improves the performance but not significantly. This shows that this feature may not be effective for wellness event detection, as we had expected, though it is widely used for public event detection. We also observed that **Modality** feature is useful for event detection. Indeed, we observed that it is able to filter out activities from wishes or general thoughts and information significantly.

On Parameter Sensitivity

An important parameter in our method is λ in Eq. (4) that determines the impact of relation amongst tasks in the learning process. A high value indicates the importance of

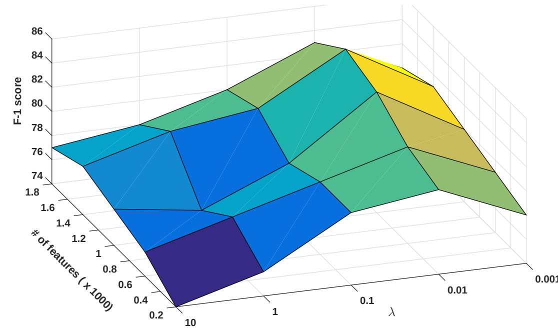

Figure 1: The impact of different parameter setting.

these relations while a low value limits the effect of relations amongst tasks in the learning process. Another important parameter is the number of selected features. Hence, we study how the performance of our model varies with λ and the number of selected features.

Figure 1 shows the performance of our model with different parameter settings which achieves the peak of 84.86% when $\lambda = 0.01$ and 1400 features was selected. The general pattern is that the performance is more sensitive to the number of selected features, and the best number of features is around 1400; furthermore, there is not a significant improvement above this point. It is worth noting that how to determine the number of features is still an open problem in data mining (Wang, Tang, and Liu 2015).

Conclusions and Future Work

Personal wellness events, in contrast to public events in social platform, are rarely discussed and deeply related to each other. In this paper, we proposed a learning framework that utilizes content information of microblogging texts as well as the relation between event categories to extract PWE from users social posts. In particular, we modeled the inter-relatedness among distinct events as a graph Laplacian which was employed as a regularization to a sparse learning framework. Thus the proposed model not only can learn task-shared and task-specific features but is also robust to noise in microblogging contents. Experimental evaluations on a real-world dataset from Twitter revealed that our proposed framework significantly outperforms the representative state-of-the-art methods.

This research begins a new research direction towards connecting social media and health informatics with many downstream applications such as personal care management, patient stratification, and personalized lifestyle planning.

Acknowledgments

This research is supported by the Singapore National Research Foundation under its International Research Centre@Singapore Funding Initiative and administered by the IDM Programme Office. Second author is also supported by, or in part by, Texas A&M University under grant number 22530700000 and 24147400000.

References

Abbar, S.; Mejova, Y.; and Weber, I. 2015. You tweet what you eat: Studying food consumption through twitter. In *SIGCHI*.

Akbari, M.; Nie, L.; and Chua, T.-S. 2015. amm: Towards adaptive ranking of multi-modal documents. *International Journal of Multimedia Information Retrieval*.

Association, A. D., et al. 2014. Standards of medical care in diabetes. *Diabetes care*.

Becker, H.; Naaman, M.; and Gravano, L. 2011. Beyond trending topics: Real-world event identification on twitter. In *ICWSM*.

Chen, X.; Pan, W.; Kwok, J. T.; and Carbonell, J. G. 2009. Accelerated gradient method for multi-task sparse learning problem. In *ICDM*.

Chen, Y.; Zhao, J.; Hu, X.; Zhang, X.; Li, Z.; and Chua, T.-S. 2013. From interest to function: Location estimation in social media. In *AAAI*.

Dos Reis, V. L., and Culotta, A. 2015. Using matched samples to estimate the effects of exercise on mental health from twitter. In *AAAI*.

Gupta, S., and Manning, C. D. 2014. Spied: Stanford pattern-based information extraction and diagnostics. *ACL*.

Hawn, C. 2009. Take two aspirin and tweet me in the morning: how twitter, facebook, and other social media are reshaping health care. *Health affairs*.

Hu, X., and Liu, H. 2012. Text analytics in social media. In *Mining text data*. Springer.

Hu, X.; Tang, J.; and Liu, H. 2014. Online social spammer detection. In *AAAI*.

Hu, F. B. 2011. Globalization of diabetes the role of diet, lifestyle, and genes. *Diabetes care*.

Kim, S., and Xing, E. P. 2009. Statistical estimation of correlated genome associations to a quantitative trait network. *PLoS genetics*.

Lee, J. L.; DeCamp, M.; Dredze, M.; Chisolm, M. S.; and Berger, Z. D. 2014. What are health-related users tweeting? a qualitative content analysis of health-related users and their messages on twitter. *JMIR*.

Li, J.; Ritter, A.; Cardie, C.; and Hovy, E. 2014. Major life event extraction from twitter based on congratulations/condolences speech acts. In *EMNLP*.

Lim, S. S.; Vos, T.; Flaxman, A. D.; et al. 2013. A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions. *The lancet*.

Meladianos, P.; Nikolentzos, G.; Rousseau, F.; Stavrakas, Y.; and Vazirgiannis, M. 2015. Degeneracy-based real-time sub-event detection in twitter stream. In *ICWSM*.

Mintz, M.; Bills, S.; Snow, R.; and Jurafsky, D. 2009. Distant supervision for relation extraction without labeled data. In *ACL*.

Nesterov, Y. 2004. Introductory lectures on convex optimization: a basic course.

Nie, F.; Huang, H.; Cai, X.; and Ding, C. H. 2010. Efficient and robust feature selection via joint $l_{2/1}$-norms minimization. In *NIPS*.

Nie, L.; Akbari, M.; Li, T.; and Chua, T.-S. 2014a. A joint local-global approach for medical terminology assignment. In *SIGIR*.

Nie, L.; Li, T.; Akbari, M.; Shen, J.; and Chua, T.-S. 2014b. Wenzher: Comprehensive vertical search for healthcare domain. In *SIGIR*.

Nie, L.; Zhao, Y.-L.; Akbari, M.; Shen, J.; and Chua, T.-S. 2015. Bridging the vocabulary gap between health seekers and healthcare knowledge. *TKDE*.

Obozinski, G.; Taskar, B.; and Jordan, M. I. 2010. Joint covariate selection and joint subspace selection for multiple classification problems. *Statistics and Computing*.

Pastors, J. G.; Warshaw, H.; Daly, A.; Franz, M.; and Kulkarni, K. 2002. The evidence for the effectiveness of medical nutrition therapy in diabetes management. *Diabetes care*.

Paul, M.; Dredze, M.; Broniatowski, D.; and Generous, N. 2015. Worldwide influenza surveillance through twitter. In *AAAI*.

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

Ritter, A.; Etzioni, O.; Clark, S.; et al. 2012. Open domain event extraction from twitter. In *SIGKDD*.

Ruvolo, P., and Eaton, E. 2014. Online multi-task learning via sparse dictionary optimization. In *AAAI*.

Shelley, K. J. 2005. Developing the american time use survey activity classification system. *Monthly Lab. Rev.*

Song, X.; Nie, L.; Zhang, L.; Akbari, M.; and Chua, T.-S. 2015. Multiple social network learning and its application in volunteerism tendency prediction. In *SIGIR*.

Tang, D.; Wei, F.; Yang, N.; Zhou, M.; Liu, T.; and Qin, B. 2014. Learning sentiment-specific word embedding for twitter sentiment classification. In *ACL*.

Teodoro, R., and Naaman, M. 2013. Fitter with twitter: Understanding personal health and fitness activity in social media. In *ICWSM*.

Thelen, M., and Riloff, E. 2002. A bootstrapping method for learning semantic lexicons using extraction pattern contexts. In *ACL*.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*.

Wang, S.; Tang, J.; and Liu, H. 2015. Embedded unsupervised feature selection. In *AAAI*.

Xu, L.; Huang, A.; Chen, J.; and Chen, E. 2015. Exploiting task-feature co-clusters in multi-task learning. In *AAAI*.

Yang, Y.; Jia, J.; Zhang, S.; Wu, B.; Chen, Q.; Li, J.; Xing, C.; and Tang, J. 2014. How do your friends on social media disclose your emotions? In *AAAI*.

Zhou, D.; Chen, L.; and He, Y. 2015. An unsupervised framework of exploring events on twitter: Filtering, extraction and categorization. In *AAAI*.