A Personal Privacy Preserving Framework: I Let You Know Who Can See What

Xuemeng Song[†], Xiang Wang[‡], Liqiang Nie[†], Xiangnan He[‡], Zhumin Chen[†] , Wei Liu[$]

[†]School of Computer Science and Technology, Shandong University

[‡]School of Computing, University of National Singapore, Singapore

[$]Tencent AI Lab

Motivation

Personal demographics

Daily activities

Relationship

⋮

Information pertaining to users themselves accounts for up to 66% of the entire user generated contents (UGCs) [1].

Motivation

graphics

Information pertainin ed contents (UGCs) [1].

Motivation

- **The default privacy settings usually make UGCs publicly accessible.**

A real story…

Video podcaster

June 2009

Looking forward to my family vacation to Saint Louis, where we would be visiting family friends for the week.

We had successfully arrived in Missouri.

Vacation at Saint Louis

Home in Arizona

Motivation

- **Users may even be unaware of the privacy leakage when they are posting on social networks, which leads to the regrettable messages [1].**

I can't believe I said that!

Privacy leakage via UGCs deserves our special attention.

Regrettable messages

[1] Sleeper, M.; Cranshaw, J.; Kelley, P. G.; Ur, B.; Acquisti, A.; Cranor, L. F.; and Sadeh, N. 2013. I read my twitter the next morning and was astonished: A conversational perspective on twitter regrets. In SIGCHI.

Related Work

Privacy

Structured Data

- User structured profiles,
- Privacy settings,
- Trajectory records…

Far too little attention has been paid to investigate users' unstructured data, whereby the data volume is larger, information is richer, and privacy issues are more prominent.

Unstructured Data

- User generated contents.

Mainly focus on training effective classifiers to predict whether the given UGC is privacy-sensitive.

Related Work

Multi-task Learning

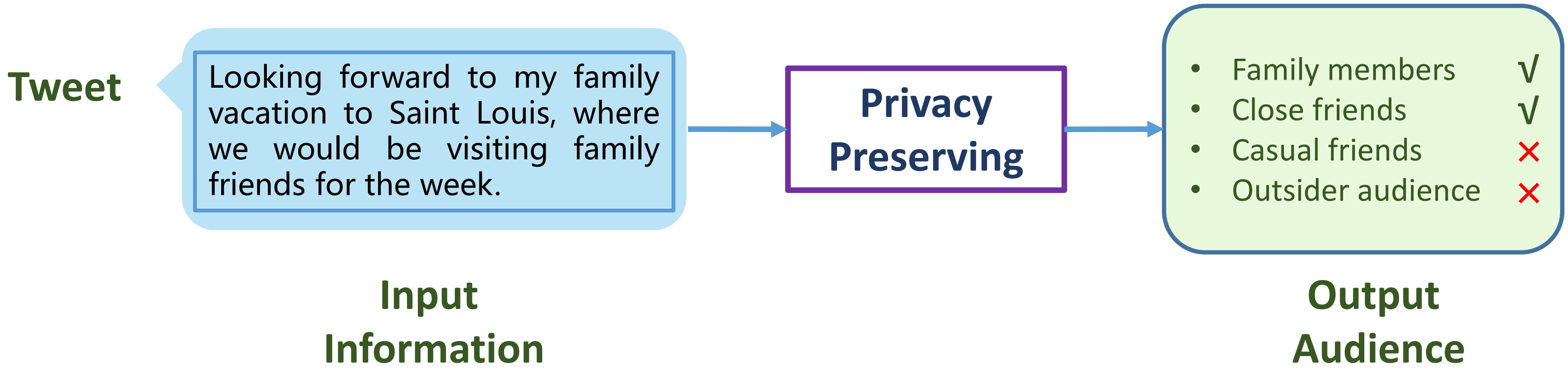
Although multi-task learning has been successfully applied to

- Social behavior prediction,
- Image annotation,
- Web search,
- …

Limited efforts have been dedicated to the privacy domain.

Task Definition

Considering that information and audience both play pivotal roles in the privacy preserving, answering the question of *Who Can See What* is essential.



Challenges

- **The personal aspects of users conveyed by their UGCs are usually not independent but related. The main challenge is how to construct and leverage the relatedness structure to boost the performance.**
- **No gold standard instruction is available to guide *Who Can See What*.**
- **The lack of benchmark dataset and the way to extract a set of privacy-oriented features.**

Framework

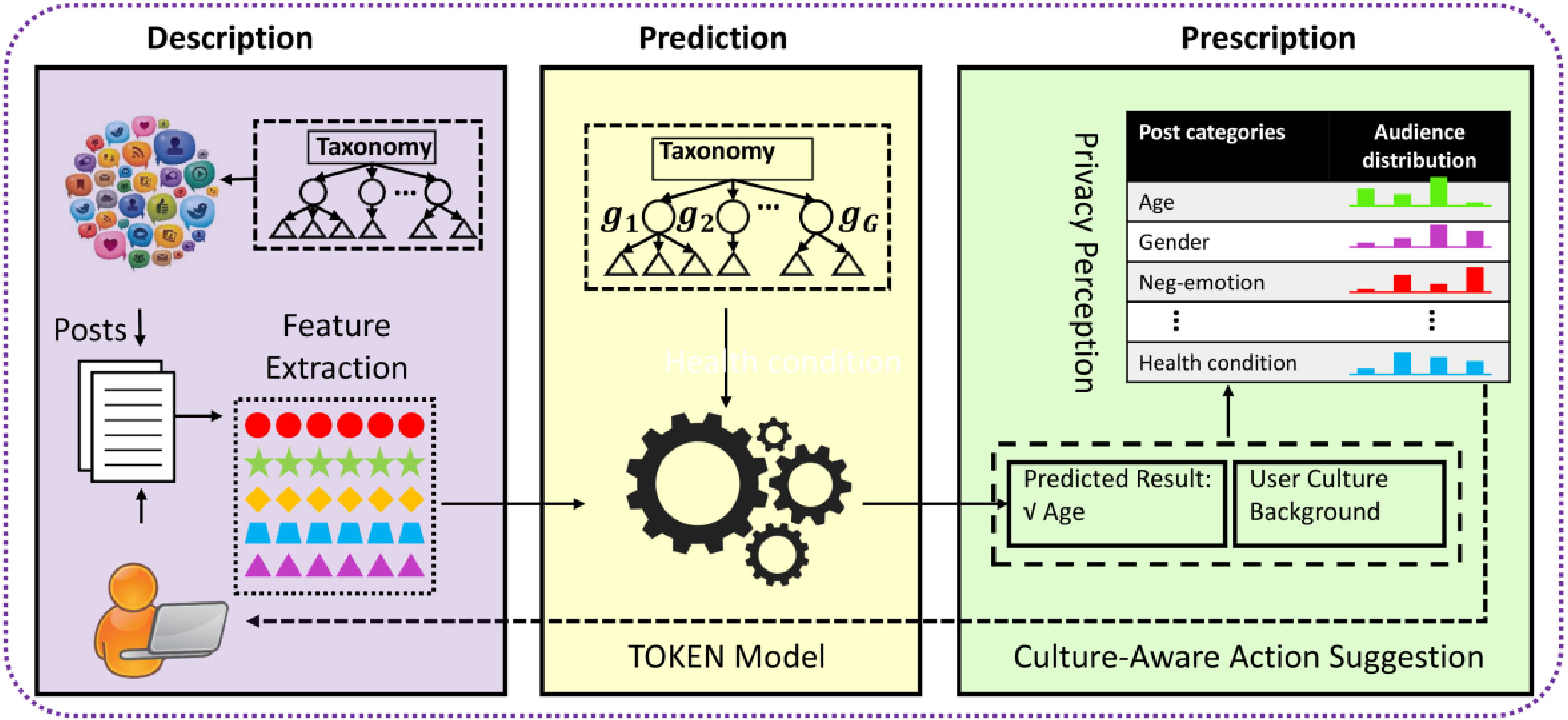


Figure 1: Illustration of the proposed scheme.

Description

Taxonomy Induction

Caliskan-Islam et al. 2014

Location

Medical

Drug

Emotion

Stereotying

Associations

Personal Attacks

Personal Details

Identifiable Information

- Coarse-grained.
- Overlook the life milestones of individuals.

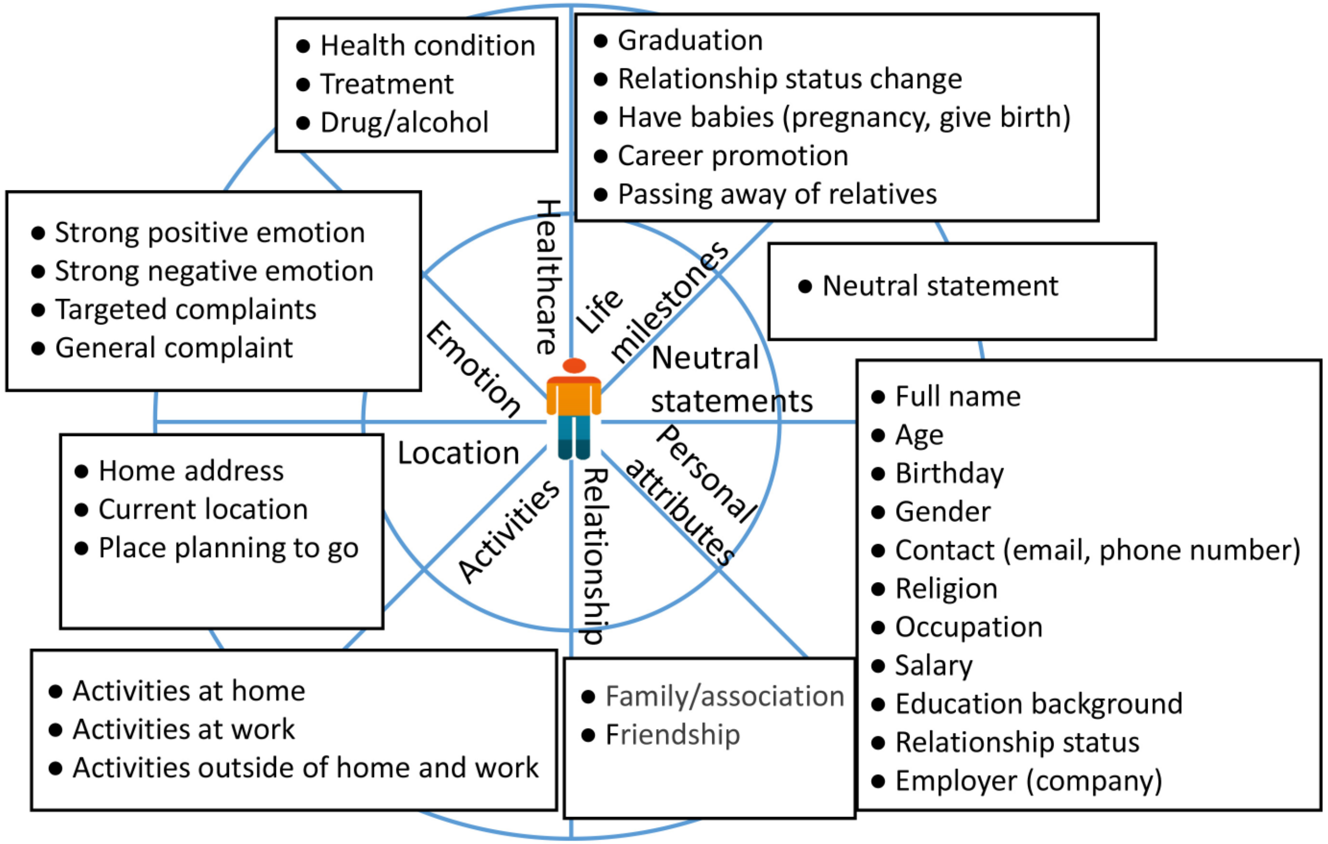


Figure 2. Illustration of our pre-defined taxonomy.

Description

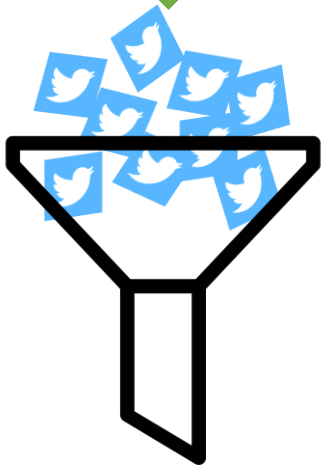
Data Collection

- **Users' tweets revealing their personal aspects are usually sparse, we hence give up the user-centric crawling policy.**

Pre-defined keywords

Twitter Search Service

269, 090 raw tweets.



11,370 tweets.

Ground Truth Construction

Three "masters" are employed for tweet annotations.

Description

Example Illustration

Table1. Examples of selected categories.

Category	Examples
Occupation	"I used to be a swimmer...now I'm a coach. And I love torturing my kids.
	"I felt more control of my work as a Teacher. "
Gender	"I seriously going to buy tacos... I am my father's daughter. "
	"The worst thing you do is piss me off while I'm on my period."
Current location	"At the Bell Performing Arts Centre for the LTS Jazz Band Concert #sweet"
	"She told the doctor tomorrow is my birthday I can't be in the hospital"
General complaint	"dude if you're going to cough every 20 seconds in the library can u leave"
	"being in a relationship is stressful i wanna take a nap"
Age	"...when I told him I'm only 24"
	"Hey @user1 its my birthday tomorrow. I am turning 12! "
Neutral statement	"Chelsea look like they got promoted last season.."
	"Do you want my home address and social security too?"

Description

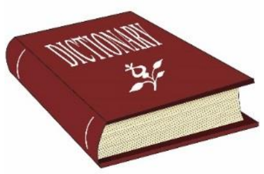
Features

- **Linguistic Inquiry Word Count (LIWC)**
- **Privacy Dictionary**
- **Sentiment Analysis**
- **Sentence2Vector**
- **Meta-features**

Description

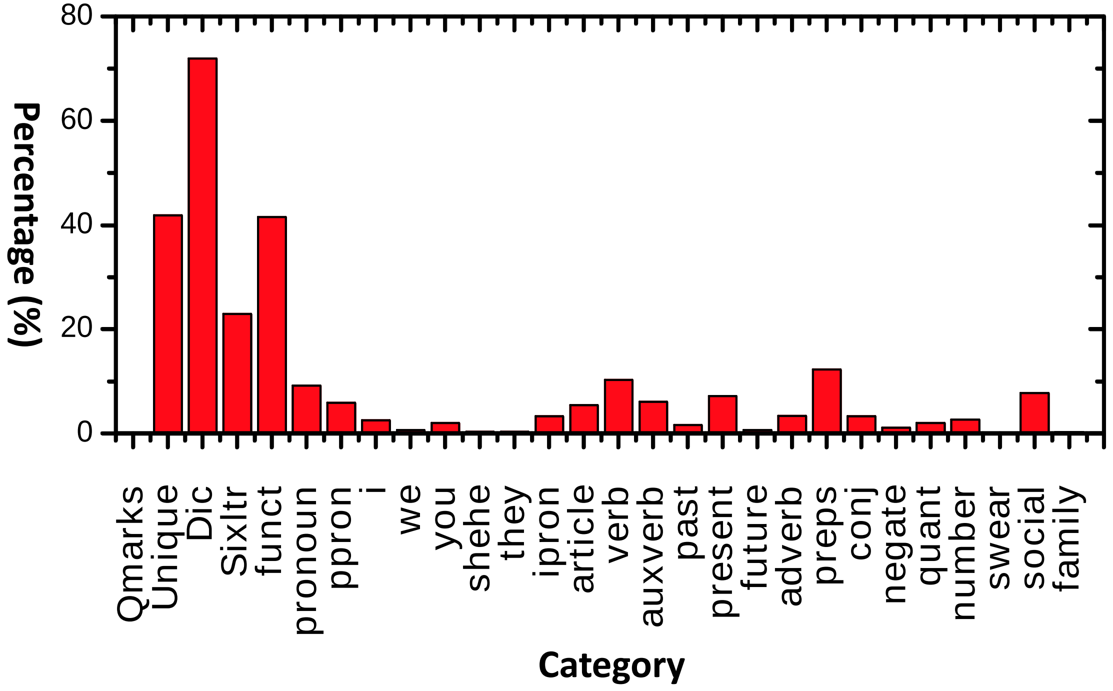
Features

- **Linguistic Inquiry Word Count (LIWC)**
- **Privacy Dictionary**
- **Sentiment Analysis**
- **Sentence2Vector**
- **Meta-features**



Dictionary

Word category



Description

Features

- Linguistic Inquiry Word Count (LIWC)
- Privacy Dictionary
- Sentiment Analysis
- Sentence2Vector
- Meta-features

Table2. Eight categories of the privacy dictionary.

Category	Explanation
OpenVisible	Represents the dialectic openness of privacy. (e.g., display, accessible.)
OutcomeState	Describes the static behavioral states and the outcomes that are served throughPrivacy. (e.g, freedom, alone.)
NormsRequisites	Encapsulates the norms, beliefs, and expectations in relation to achieving privacy. (e.g., consent, respect.)
Restriction	Expresses the closed, restrictive, and regulatory behaviors employed in maintaining privacy. (e.g., lock, exclude.)
NegativePrivacy	Captures the antecedents and consequences of privacy violations. (e.g., troubled, interfere.)
Intimacy	Portrays and measures different facets of small-group privacy. (e.g., trust, friendship.)
PrivateSecret	Expresses the “content” of privacy. (e.g., secret, data.)
Law	Describes legal definitions of privacy. (e.g., offence.)

Description

Features

- Linguistic Inquiry Word Count (LIWC)
- Privacy Dictionary
- Sentiment Analysis
- Sentence2Vector
- Meta-features

Personal Aspects

- 😀 Graduation
- 😀 Have babies
- 😀 Career promotion
- 😟 Medical treatment
- 😟 Passing away of relatives

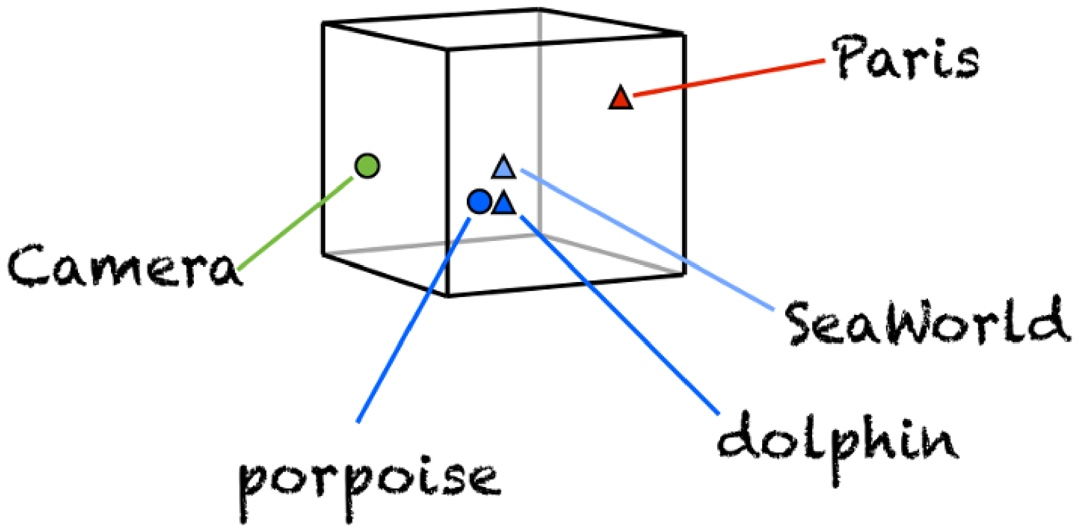
Stanford *NLP* sentiment classifier

Description

Features

- Linguistic Inquiry Word Count (LIWC)
- Privacy Dictionary
- Sentiment Analysis
- Sentence2Vector
- Meta-features

Developed based on Word2Vector . Given a tweet, Word2Vector would project it to a fixed dimensional space, where similar words are encoded spatially.



Description

Features

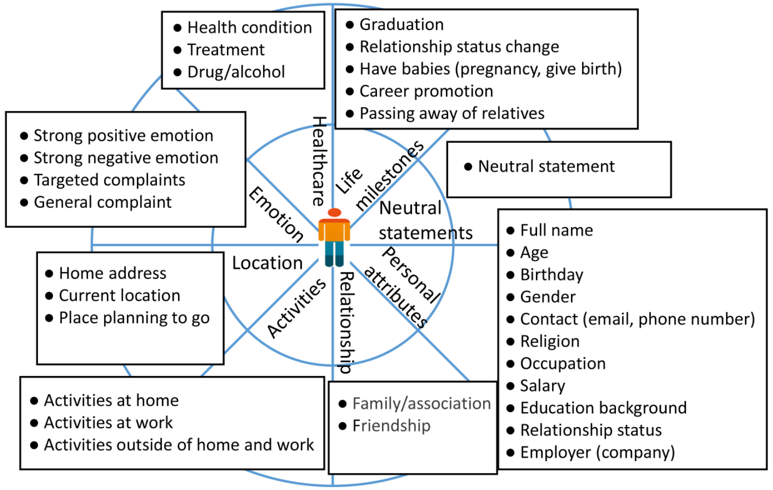
- Linguistic Inquiry Word Count (LIWC)
- Privacy Dictionary
- Sentiment Analysis
- Sentence2Vector
- Meta-features

- The presence of hashtags, slang words, images, emojis, user mentions.
- Timestamp (hour).

Eg. Happy Birthday @_slimdawg I love and miss you so much, you'll always be my best friend
7:24 PM - 1 Dec 2015

Eg. Getting drunk in a restaurant http://service.rss2twi.com/link/BeerReddit/?post_id=17561480
8:10 PM - 1 Dec 2015

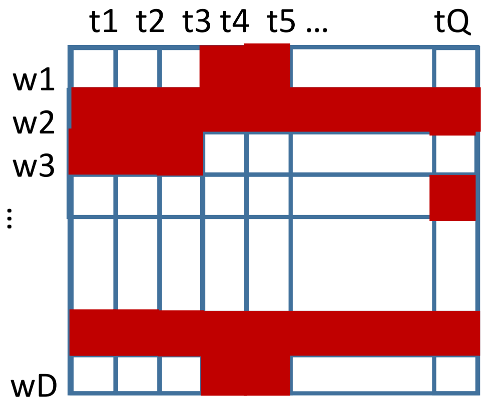
Prediction



Traditional Multi-task Feature Learning with $l_{2,1}$-norm

G groups; Q tasks; D-dimensional features.

$$\Gamma = L(\mathbf{X}, \mathbf{Y}; \mathbf{W}) + \frac{\beta}{2} \sum_{d=1}^{D} \|\mathbf{w}^d\|,$$



All tasks are related and share the common set of relevant features.

But...

It is not realistic...

Prediction

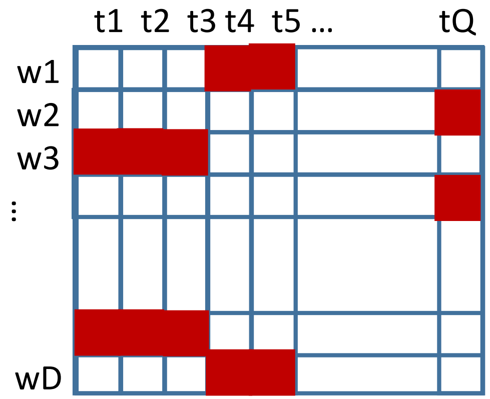
➢Group-sharing features learning

G groups; Q tasks; D-dimensional features.

$$\Gamma = L(\mathbf{X}, \mathbf{Y}; \mathbf{W}) + \frac{\beta}{2} \sum_{d=1}^{D} \|\mathbf{w}^d\|,$$

$$\Gamma = L(\mathbf{X}, \mathbf{Y}; \mathbf{W}) + \frac{\beta}{2} \sum_{g=1}^{G} \sum_{d=1}^{D} \|(\mathbf{W}\mathbf{V}_g)^d\|.$$

Group indicator matrix



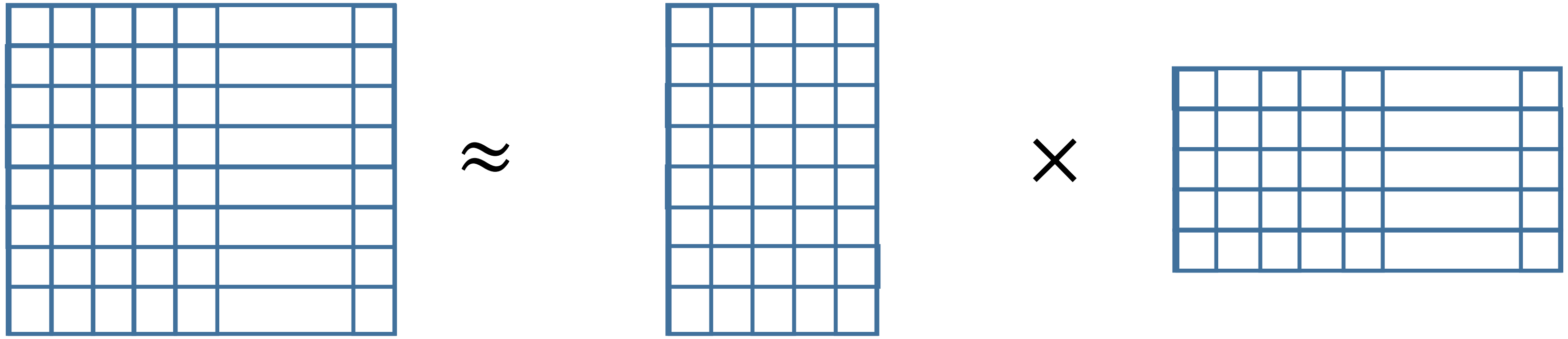
Considering that Low level features maybe not robust...

Prediction

➢High-level latent features

G groups; Q tasks; D-dimensional features.

Original (low-level) space | **Latent (semantic) space** | **Semantic representation**



$\mathbf{W} \in \boldsymbol{R}^{D*Q}$ $\approx$ $\mathbf{L} \in \boldsymbol{R}^{D*J}$ $\times$ $\mathbf{S} \in \boldsymbol{R}^{J*Q}$

J ≤ D

J is the feature dimension of latent space.

$$\Gamma = L(\mathbf{W}) + \frac{\beta}{2} \sum_{g=1}^{G} \sum_{d=1}^{D} \|(\mathbf{W}\mathbf{V}_g)^d\| \Rightarrow \Gamma = L(\mathbf{L}, \mathbf{S}) + \frac{\beta}{2} \sum_{g=1}^{G} \sum_{j=1}^{J} \|(\mathbf{S}\mathbf{V}_g)^j\|$$

Prediction

➢ **laTent grOup multi-task lEarniNg (TOKEN)**

G groups; Q tasks; D-dimensional features.

Individual-specific feature learning

Avoid overfitting

$$\min_{\mathbf{L},\mathbf{S}} L(\mathbf{L},\mathbf{S}) + \frac{\beta}{2}\sum_{g=1}^{G}\sum_{j=1}^{J}\left\|(\mathbf{S}\mathbf{V}_g)^j\right\| + \frac{\gamma}{2}\left\|\mathbf{S}\right\|_1 + \frac{\mu}{2}\left\|\mathbf{L}\right\|_F^2$$

Loss function

group-sharing feature learning

Prescription

➢Guideline Construction

- Conduct a **user study** via AMT to build guidelines regrading disclosure norms in different circles.
- Launch a cross-cultural study within two distinct areas: the U.S. and Asia12, where for each area, we hired 200 subjects.
- **Questionnaire**: a series of questions of whether he/she feels comfortable to share the given personal aspect to four social circles: ***Family members, Close Friends, Casual Friends*** and ***Outsider Audience***.
- Get **two tables of guidelines**, showing the privacy perception of users from the U.S. and Asia, respectively.

AMT

Questionnaire

Prescription

➢ **Action Suggestion**

- Based on the prediction component, we can infer which personal aspects have been leaked from the given UGC.

- Once the privacy leakage is detected, we can remind users of what has been uncovered and accordingly recommend the appropriate UGC-level privacy settings.

Experiment

Baselines

- **SVM**: This baseline simply learns each task individually. We chose the learning formulation with the kernel of radial-basis function.
- **MTL_Lasso**: The second baseline is the multi-task learning with Lasso [42]. This model also does not take advantage of prior knowledge about tasks relatedness .
- **MTFL**: The third baseline is the multi-task feature learning [2], which takes advantage of the group lasso to jointly learn features for different tasks.
- **GO-MTL (without taxonomy)**: The fourth baseline is the grouping and overlap in multi-task learning proposed in [27]. This model does not leverage the prior knowledge of task relations, as there is no taxonomy constructed to guide the learning.

Experimental Results

- **Evaluation of Description**

Table 3. Performance comparison of our model trained with different feature configurations. (%)

Features	S@K			P@K			p-value
	S@1	S@3	S@5	P@1	P@3	P@5	S@5
Privacy dictionary	8.56 ± 0.73	18.38 ± 0.78	54.26 ± 1.54	8.56 ± 0.73	6.33 ± 0.25	11.28 ± 0.36	5.9e−22
Sentiment	30.48 ± 1.51	52.23 ± 1.09	63.10 ± 1.28	30.48 ± 1.51	17.44 ± 0.36	13.32 ± 0.25	1.6e−20
Meta-features	30.31 ± 1.48	52.28 ± 1.08	63.12 ± 1.23	30.31 ± 1.48	17.38 ± 0.49	13.10 ± 0.65	9.9e−21
Sentence2Vector	33.29 ± 1.77	59.06 ± 0.97	70.91 ± 0.54	33.29 ± 1.77	20.66 ± 0.34	15.54 ± 0.17	2.0e−21
LIWC	37.13 ± 2.45	67.98 ± 1.50	78.65 ± 1.42	37.13 ± 2.45	24.72 ± 0.70	17.44 ± 0.54	3.1e−10
Total	44.37 ± 1.33	74.67 ± 1.38	84.66 ± 0.59	44.37 ± 1.33	28.42 ± 0.57	19.86 ± 0.32	-

Experimental Results

- **Evaluation of Description**

Table 3. Performance comparison of our model trained with different feature configurations. (%)

Features	S@K			P@K			p-value
	$S@1$	$S@3$	$S@5$	$P@1$	$P@3$	$P@5$	S@5
Privacy dictionary	8.56 ± 0.73	18.38 ± 0.78	54.26 ± 1.54	8.56 ± 0.73	6.33 ± 0.25	11.28 ± 0.36	5.9e−22
Sentiment	30.48 ± 1.51	52.23 ± 1.09	63.10 ± 1.28	30.48 ± 1.51	17.44 ± 0.36	13.32 ± 0.25	1.6e−20
Meta-features	30.31 ± 1.48	52.28 ± 1.08	63.12 ± 1.23	30.31 ± 1.48	17.38 ± 0.49	13.10 ± 0.65	9.9e−21
Sentence2Vector	33.29 ± 1.77	59.06 ± 0.97	70.91 ± 0.54	33.29 ± 1.77	20.66 ± 0.34	15.54 ± 0.17	2.0e−21
LIWC	37.13 ± 2.45	67.98 ± 1.50	78.65 ± 1.42	37.13 ± 2.45	24.72 ± 0.70	17.44 ± 0.54	3.1e−10
Total	44.37 ± 1.33	74.67 ± 1.38	84.66 ± 0.59	44.37 ± 1.33	28.42 ± 0.57	19.86 ± 0.32	-

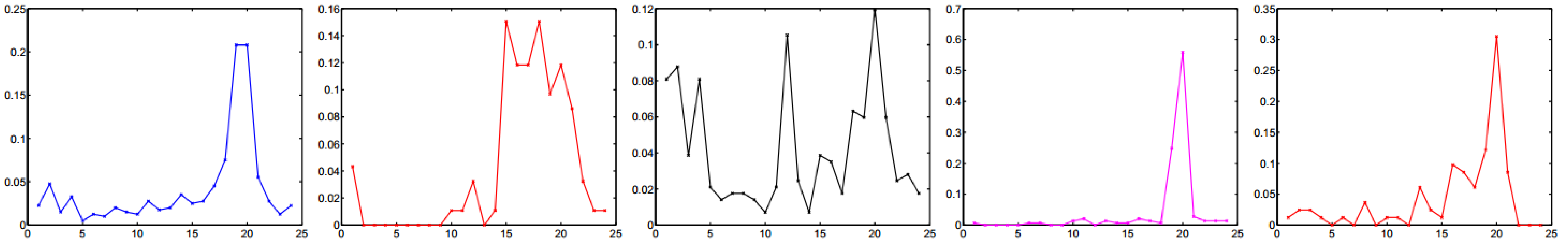
LIWC

Content categories: 'home', 'job', 'social'…
Style categories: pronouns ('first', 'second', 'third'), verb tense ('past', 'present', 'future')…

self- or other-references and temporal hints

Experimental Results

➢ Evaluation of Description



(a) Activities Outside (b) Activities at Work (c) Activities at Home (d) Drug/Alcohol (e) Status Change

Figure 3. Illustration of temporal patterns regarding personal aspects. X axis: the timeline (by hour); Y axis: the distribution of tweets.

Experimental Results

➢ **Evaluation of Prediction**

Table 4. Performance comparison between our TOKEN model and the baselines in S@K and P@K (%).

Methods	S@K			P@K			p-value
	S@1	S@3	S@5	P@1	P@3	P@5	S@5
SVM	2.65 ± 1.09	52.15 ± 4.25	72.01 ± 1.28	2.65 ± 1.09	17.80 ± 2.03	16.53 ± 0.52	2.3e−16
MTL_Lasso	43.99 ± 1.18	73.02 ± 1.30	82.26 ± 0.83	43.99 ± 1.18	27.35 ± 0.56	19.34 ± 0.26	6.9e−7
MTFL	43.75 ± 2.03	73.98 ± 1.03	83.69 ± 0.68	43.75 ± 2.03	27.63 ± 0.51	19.70 ± 0.28	3.1e−3
GO-MTL	43.92 ± 1.29	73.93 ± 1.15	83.45 ± 0.94	43.92 ± 1.29	27.25 ± 0.45	19.40 ± 0.31	2.9e−3
TOKEN	44.37 ± 1.33	74.67 ± 1.38	84.66 ± 0.59	44.37 ± 1.33	28.42 ± 0.57	19.86 ± 0.32	-

Experimental Results

➢ Evaluation of Prescription Analysis

On the Cultural Privacy Perception

Table 5: The eight categories with the most different privacy perceptions between the U.S. and Asia. The percentage of subjects who feel comfortable to share the given personal aspect to each social circle. FA: Family Member; CL: Close Friends; CA: Casual Friends; OU: Outsider Audience.

Categories	the U.S.				Asia			
	FA	CL	CA	OU	FA	CL	CA	OU
emotion: positive emotion	95.0%	97.5%	83.0%	54.0%	75.5%	86.5%	44.5%	21.0%
emotion: negative emotion	88.5%	93.5%	59.5%	36.5%	49.0%	77.5%	31.0%	20.0%
personal attributes: gender	95.5%	96.0%	84.5%	63.5%	75.5%	76.5%	53.0%	32.5%
emotion: general complaints	92.0%	94.0%	83.5%	59.5%	67.0%	79.0%	52.0%	32.0%
personal attributes: age	98.5%	96.0%	74.5%	40.0%	89.5%	79.0%	38.0%	16.0%
activity: activities at home	95.0%	93.0%	61.5%	35.0%	79.0%	68.5%	33.5%	13.0%
neutral statements	98.0%	96.5%	94.0%	85.5%	75.0%	81.0%	70.5%	65.5%

Experimental Results

➢ Evaluation of Prescription Analysis

On the Cultural Privacy Perception

Table 6: The eight categories with the most similar privacy perceptions between the U.S. and Asia. The percentage of subjects who feel comfortable to share the given personal aspect to each social circle. FA: Family Member; CL: Close Friends; CA: Casual Friends; OU: Outsider Audience.

Categories	the U.S.				Asia			
	FA	CL	CA	OU	FA	CL	CA	OU
healthcare: treatments	96.0%	76.5%	18.5%	5.0%	88.0%	65.5%	14.5%	7.5%
healthcare: health conditions	98.0%	71.0%	17.5%	7.0%	85.0%	65.5%	19.5%	7.5%
life milestones: passing away	95.5%	86.5%	35.0%	12.0%	88.0%	74.5%	31.0%	7.5%
emotion: specific complaints	53.5%	78.0%	28.0%	17.5%	36.5%	68.0%	28.0%	19.0%
location: home address	95.5%	71.0%	5.0%	3.0%	80.5%	73.0%	18.0%	6.0%
location: current location	94.5%	87.5%	31.5%	9.0%	75.0%	77.5%	35.0%	11.5%
personal attributes: contact	95.5%	87.0%	18.5%	3.0%	77.5%	80.5%	27.5%	10.5%
location: places planning to go	95.0%	91.5%	51.0%	21.5%	77.0%	87.0%	39.0%	13.5%

Conclusion

We study the problem of privacy preserving by presenting a scheme, consisting of three components: description, prediction and prescription.

- **As to description, we build a comprehensive taxonomy, construct a benchmark dataset, and develop a set of privacy-oriented features.**
- **Regarding prediction, we propose a taxonomy-guided multi-task learning model to categorize social posts, which is able to learn both group-sharing and aspect-specific features simultaneously.**
- **In terms of prescription, we construct cross-culture guidelines regarding the user's information disclosure norms based on the crowd intelligence via AMT.**

Future Work

- Currently, we only explore the simple linear mapping to model the prediction component. However, the complicated prediction mapping may lie in the non-linear space.
- We plan to extend our work towards applying the more advanced neural networks in our context.

Thank You

Back Up

Experimental Results

- **Evaluation of Description**

Table 3. Performance comparison of our model trained with different feature configurations. (%)

Features	S@K			P@K			p-value
	S@1	S@3	S@5	P@1	P@3	P@5	S@5
Privacy dictionary	8.56 ± 0.73	18.38 ± 0.78	54.26 ± 1.54	8.56 ± 0.73	6.33 ± 0.25	11.28 ± 0.36	5.9e−22
Sentiment	30.48 ± 1.51	52.23 ± 1.09	63.10 ± 1.28	30.48 ± 1.51	17.44 ± 0.36	13.32 ± 0.25	1.6e−20
Meta-features	30.31 ± 1.48	52.28 ± 1.08	63.12 ± 1.23	30.31 ± 1.48	17.38 ± 0.49	13.10 ± 0.65	9.9e−21
Sentence2Vector	33.29 ± 1.77	59.06 ± 0.97	70.91 ± 0.54	33.29 ± 1.77	20.66 ± 0.34	15.54 ± 0.17	2.0e−21
LIWC	37.13 ± 2.45	67.98 ± 1.50	78.65 ± 1.42	37.13 ± 2.45	24.72 ± 0.70	17.44 ± 0.54	3.1e−10
Total	44.37 ± 1.33	74.67 ± 1.38	84.66 ± 0.59	44.37 ± 1.33	28.42 ± 0.57	19.86 ± 0.32	-

Privacy_dictionary

Law, OpenVisible, OutcomeState, NormsRequisites, Restriction, NegativePrivacy, Intimacy, and PrivateSecret

Formal/ professional
Small-scale