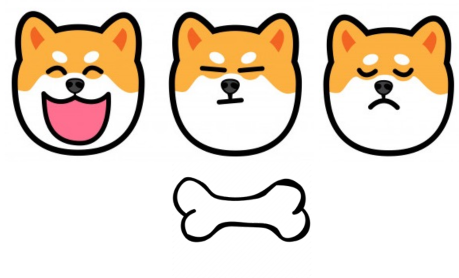


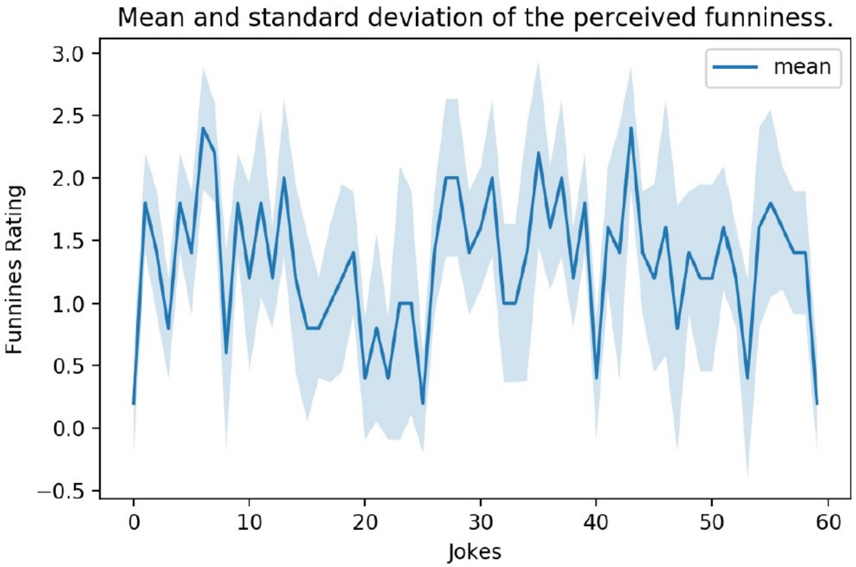
Federated Learning for Personalized Humor Recognition

Xu Guo, Han Yu, Boyang Li, Hao Wang,Pengwei Xing, Siwei Feng, Zaiqing Nie and Chunyan Miao

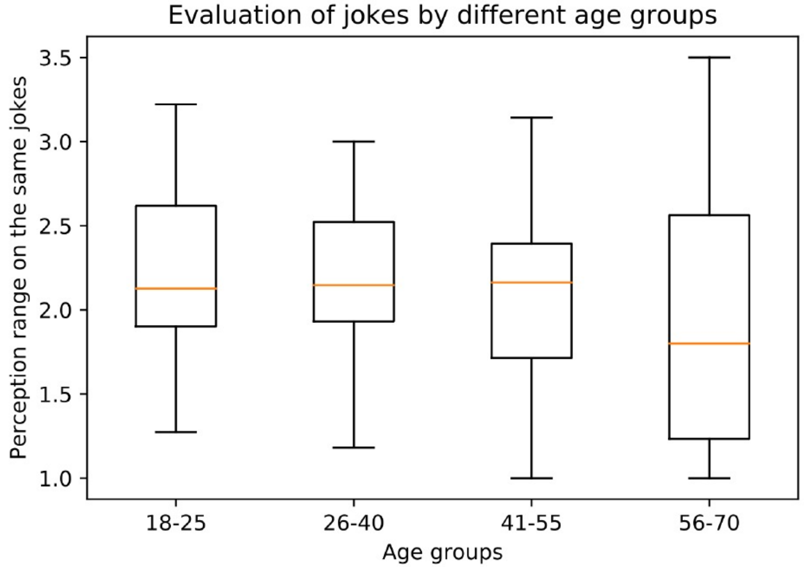
ACM Transactions on Intelligent Systems and Technology (TIST) 13.4 (2022): 1-18.

Background and Motivation

- Human perception of a joke is highly *subjective*
- Observation: every joke can be perceived differently by different people



(a)

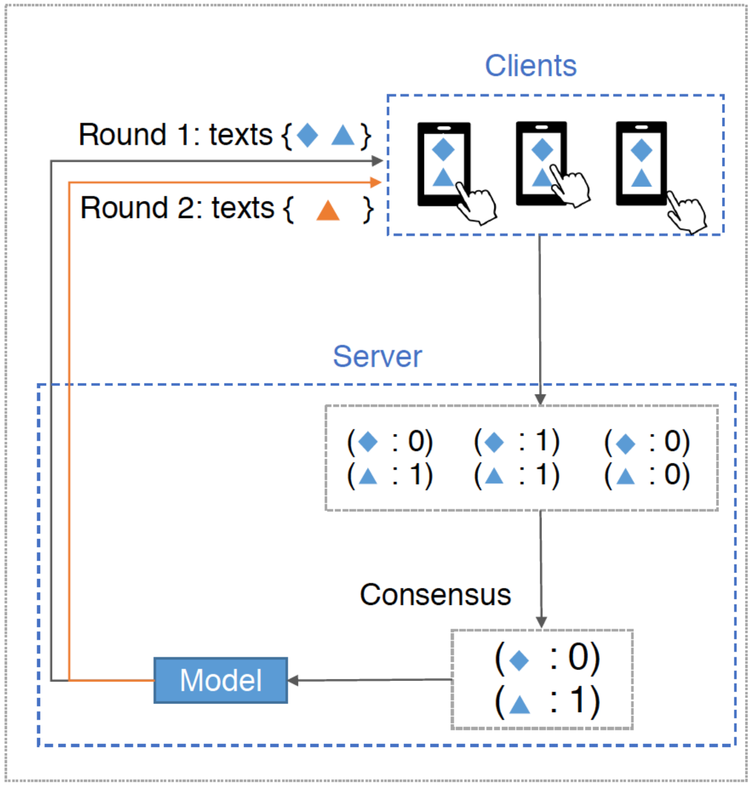


(b)

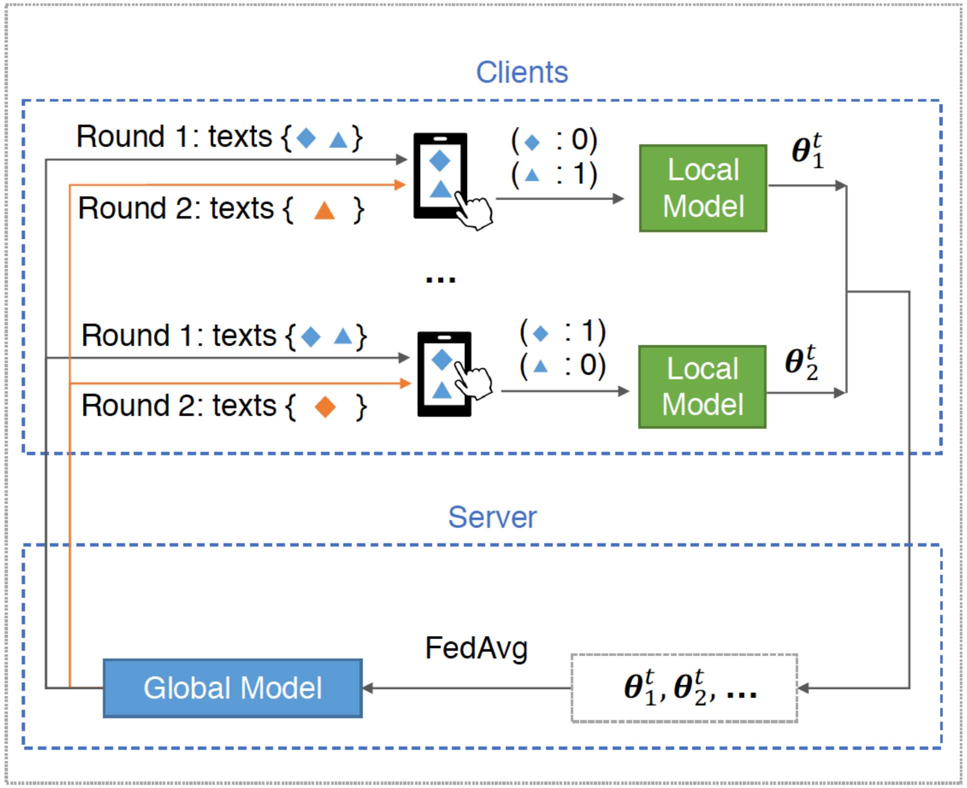
Data from "President Vows to Cut ~~Taxes~~ Hair": Dataset and Analysis of Creative Text Editing for Humorous Headlines. 2019. In NAACL.

Personalized humor recognition through federated learning

Traditional



Ours



Weight-tying Federated Updates

Update θ_i^t in regular gradient descent:

$$\theta_i^{t,0} = \theta^t$$

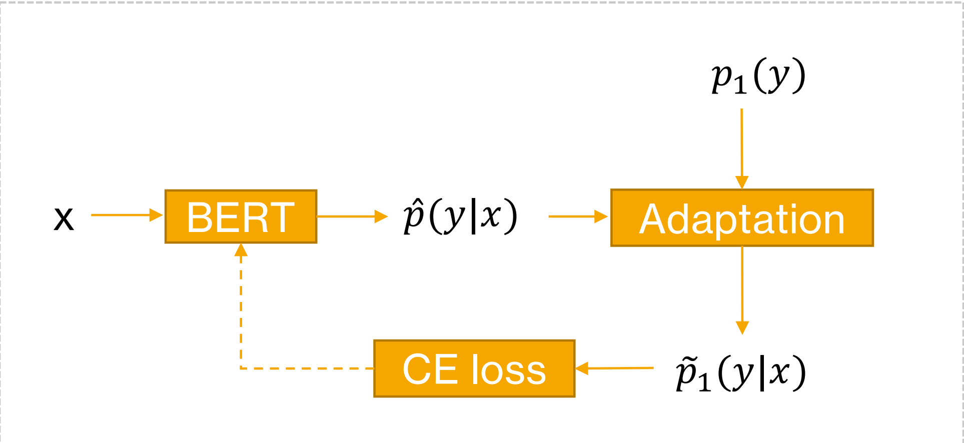
$$\theta_i^{t,K} = \arg\min_{\theta_i^t} L(\theta_i^t; \alpha_i) \tag{5}$$

Weight tying is achieved using federated averaging

$$\theta^{t+1} \leftarrow \frac{1}{m} \sum_{i=1}^{m} \theta_i^{t,K} \tag{6}$$

Repeat

Local Adaptation with FedHumor



- Adaptation
  - $\tilde{p}_i(y|x, \alpha_i) = \mathrm{Softmax}\left(\frac{\hat{p}_i(y|x)}{p_i(y)^{\beta_i}}\right)$ (1)
- $p_i(y)$ is the local empirical label distribution
  - $p_i(y) = \frac{1}{|D|}\sum_j \mathbb{I}\,(y_{i,j} = funny)$ (2)
- β_i is a hyperparameter determined on validation set.

Objective function: $L_i(\theta_i^t) = -\,y_i \log \tilde{p}_i(y|x) + \lambda||\theta_i^t||_2^2$ (3)

Comparison of Different Training Strategies

- Data:
  - Differently and independently distributed
- Training Approach:
  - AGG: *aggregate* all the labelled data and train on a centralized setting.
  - INDV: *individually* train a model for each user.
  - FED: using *federated averaging* to tie weights.
- Testing scenarios:
  - Group 1: a group of 3 users with unique preferences
  - Group 2: a group of 18 users with unique preferences

Hypothesis: Federated learning creates an ensemble model

Table 2: (Average) Test performance

		Precision	Recall	F_1 score
Group 1	AGG	<u>58.59</u>	54.89	41.66
	INDV	56.30	<u>55.32</u>	<u>53.52</u>
	FED	**60.03**	**65.57**	**55.61**
Group 2	AGG	57.40	51.25	33.05
	INDV	<u>58.14</u>	<u>55.61</u>	<u>53.03</u>
	FED	**61.67**	**66.62**	**57.48**

Comparison of Different Humor Recognition Models

- Without pretrained language model
  - DV-LR: Document vectors + Logistic Regression
  - WV-RF: word2vector + Random Forest Classifier
  - WV-CNN-HN: word2vector + CNN + Highway + fully connected classifier
- With pretrained language model
  - BERT-FZ/FT: BERT base version with pretrained weights *Freezed* or *FineTuned*
  - BERT-L/C/M: BERT Large or Cased or Multilingual Versions.
  - ALBERT: Faster BERT
- With pretrained language model and federated training strategy
  - FedHumor: BERT base + Federated Training

Table 3: Test performance (macro-averaged)

	Precision	Recall	F_1 score
DV-LR	53.69	53.67	53.64
WV-RF	56.70	56.10	55.20
WV-CNN-HN	56.20	54.70	51.90
BERT-FZ	54.15	53.71	52.53
BERT-FT	64.91	64.88	64.87
BERT-L	64.48	64.48	64.47
BERT-C	62.69	62.65	62.62
BERT-M	62.11	62.08	62.06
ALBERT	61.06	61.05	61.04
FedHumor	**66.60**	**66.56**	**66.53**

Thank You

- Presenter: Xu GUO (xu008@e.ntu.edu.sg)

Experiment – Dataset

- We use the SemEval-2020 shared Task 7 - assessing the funniness of edited news headlines – for experiments.
- The original dataset contains the average ratings from 5 human annotators using a value in the range [0, 1, 2, 3].

Table 1: Statistics of the public dataset

	Train	Validation	Test
Number of samples	9,652	2,419	3,024
Average Rating	0.936	0.935	0.940
Minimum Rating	0.000	0.000	0.000
Maximum Rating	3.000	3.000	2.800

Synthetic Data Generation

- Sort the jokes by their original average ratings.
- A user has only one humor preference α_i.
- α_i is defined as a transition point in the funniness interval.