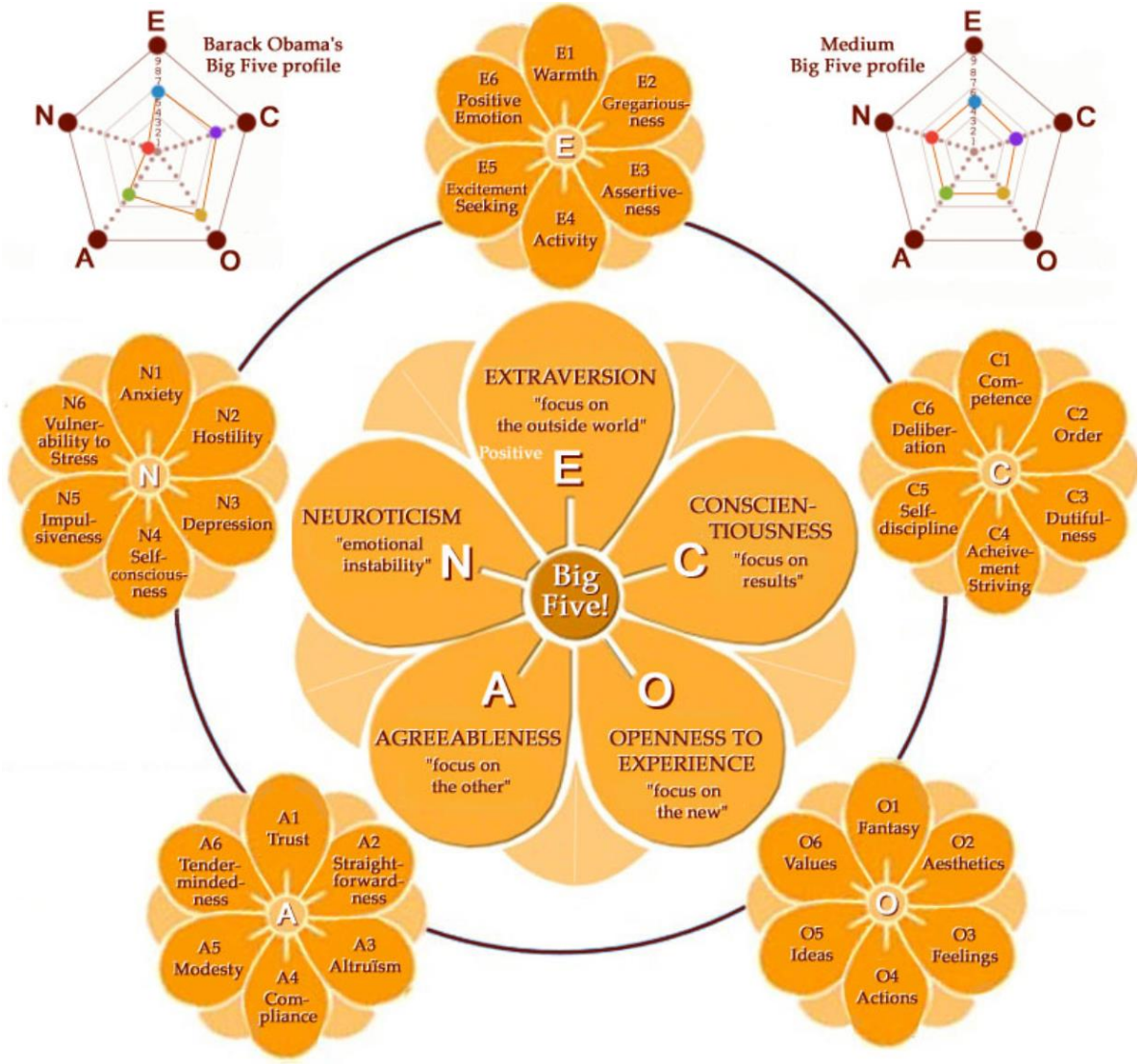
Beyond the Words: Predicting User Personality from Heterogeneous Information

张富峥
微软亚洲研究院

https://www.microsoft.com/en-us/research/people/fuzzhang/
fuzzhang@microsoft.com

Microsoft

Big 5 Personality



Previous Work

- Psychology (questionnaire-based) approaches
  - Surveys or interview-based approaches
  - Rely heavily on retrospective self-reports
  - Time costs, money costs, data granularity and etc.

- Data-driven methods
  - Texts, such a Closed vocabulary, *Linguistic Inquiry and Word Count(LIWC)*

    Open vocabulary*, Topic Modeling(LDA)*
  - Single feature in user digital traces, such as

    Hash-tags and Facebook likes *(Mypersonality)*

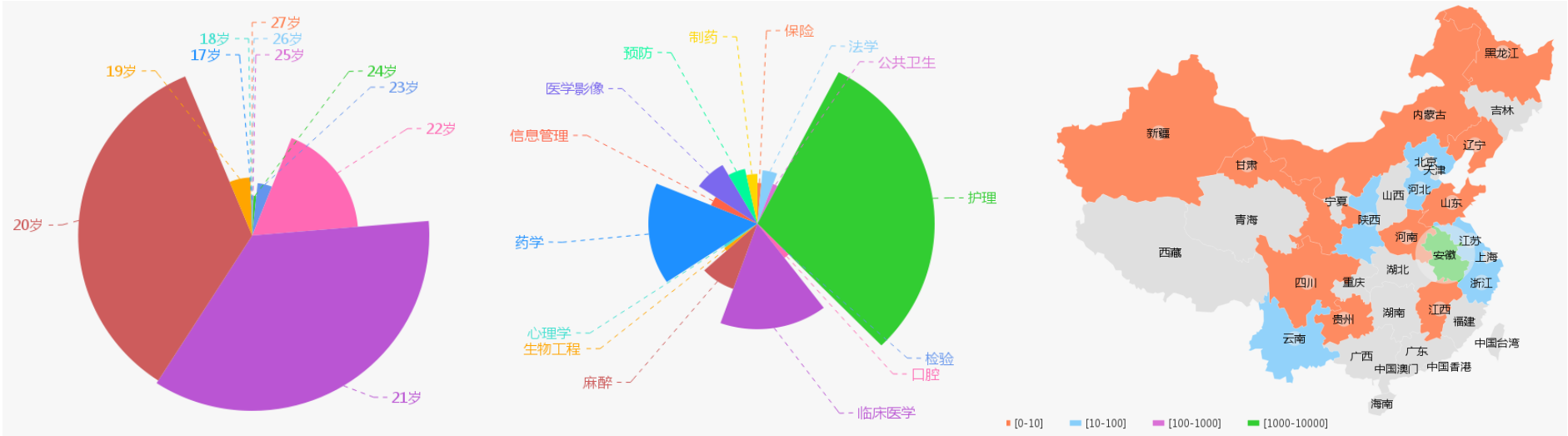
Heterogeneous information on social media

- ✓ Text
- ✓ Image
- ✓ User Interaction
- ✓ Emoticon

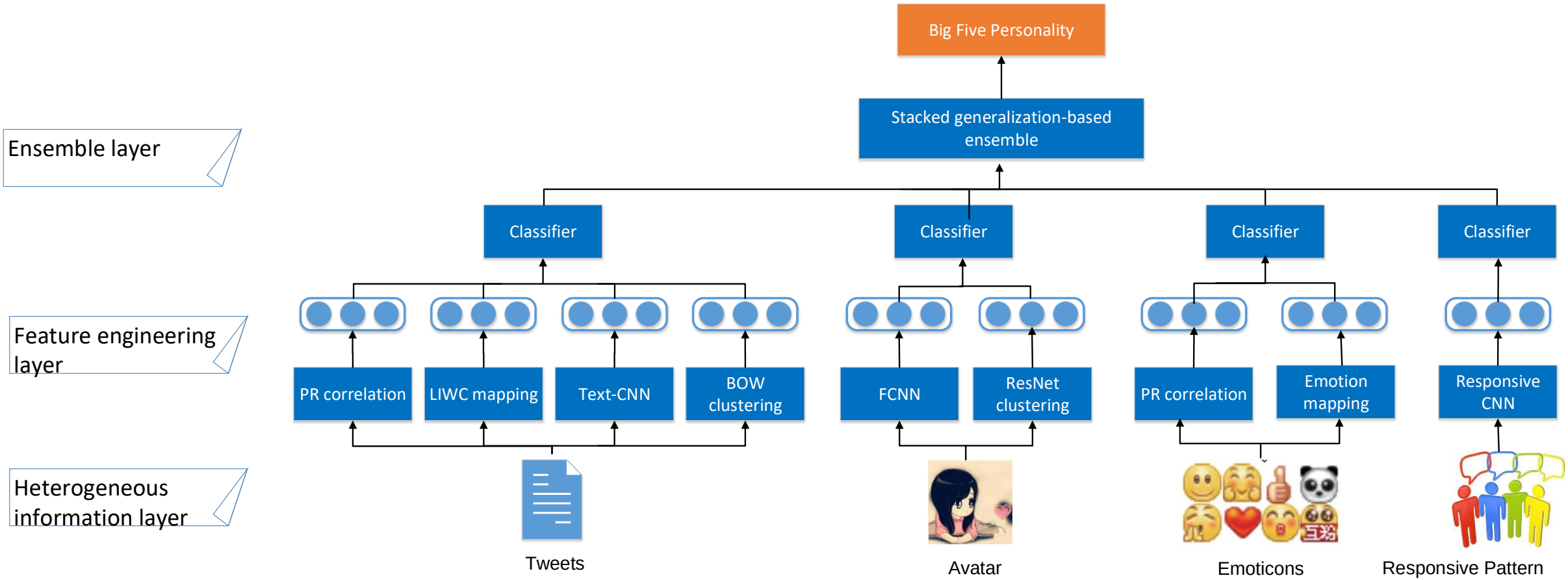
Personality Data

- 3,162 users in a medical school in Anhui,
- Test Big Five Personality with 44-item questionnaire

  Extraversion|Agreeableness|Conscientiousness|Neuroticism|Openness
- Age: average, 20.84; the majority, from 20 to 22
- Gender: Female users are the majority
- Major: nursing (n = 524), clinical medicine (n = 365) and pharmaceutics (n = 342)
- Ethnicity: Han Chinese and Hui minority
- Region: Anhui, Zhejiang, and Jiangsu



Framework: Heterogeneous Information Ensemble (HIE)



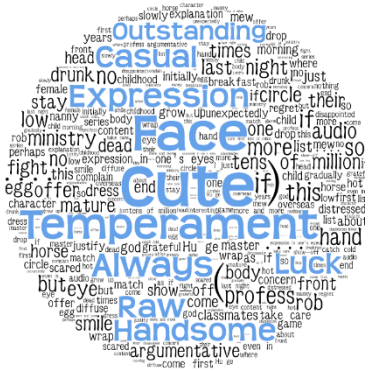
Tweets

- Text matters!
  - High in Conscientiousness:

  Formal words in Journalese,

  such as “era” and “society”
  - Low in Conscientiousness:

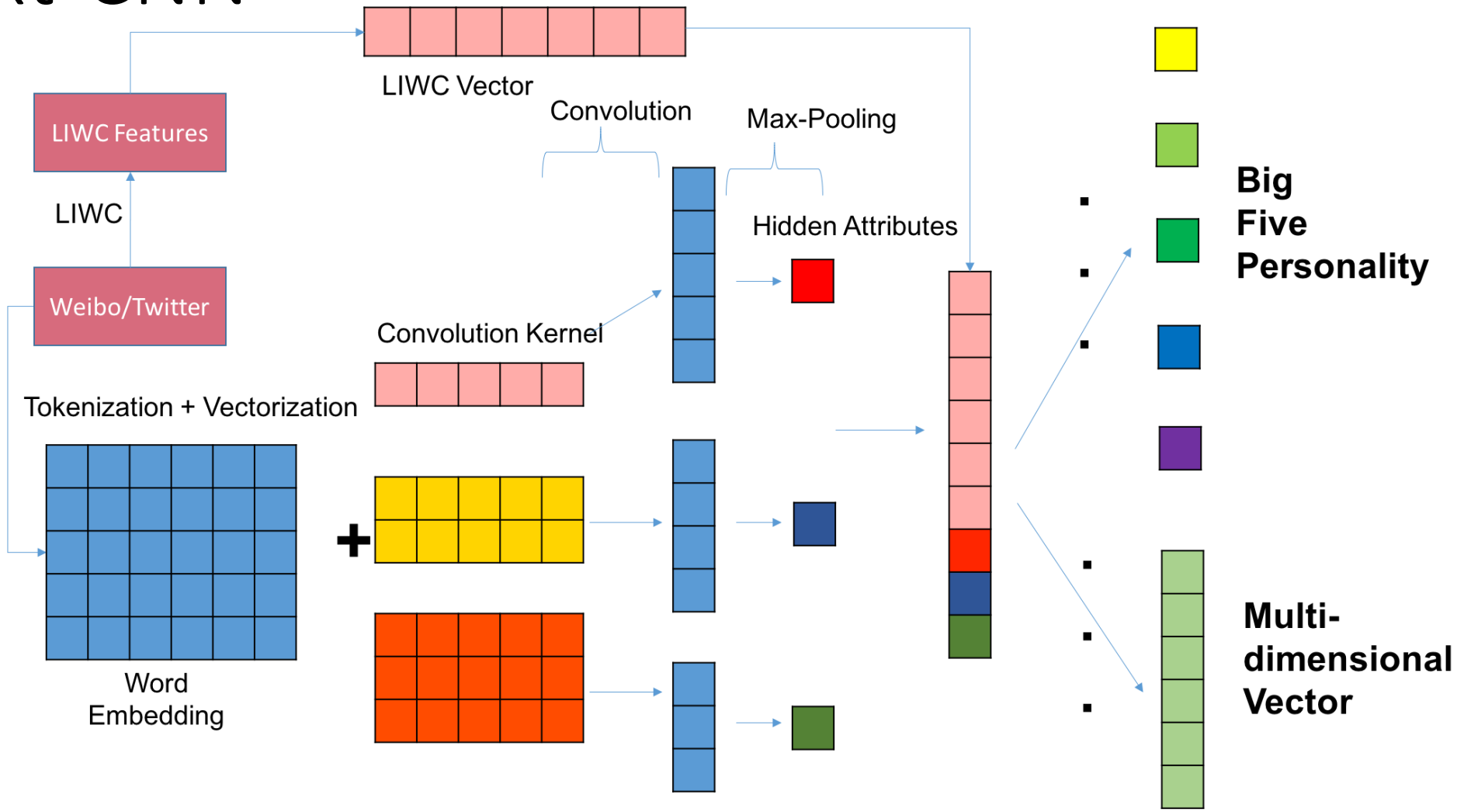
  Informal words, such as single characters or typical cyberwords.



- Pearson Correlation
  - Top 2,000 correlated words.

- Clustering
  - Top 1,500 Chinese words and all punctuations in word-embedding format.
  - K-means to cluster

Text-CNN


LIWC Vector
LIWC Features
LIWC
Weibo/Twitter
Convolution
Max-Pooling
Hidden Attributes
Convolution Kernel
Tokenization + Vectorization
Word
Embedding
Big
Five
Personality
Multi-
dimensional
Vector

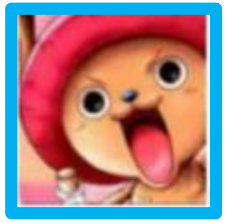
Take their avatars as an instance.

Introverts

Introverts tend to cover their face or show side face.

Extroverts are more likely to use cartoon avatars.

Extroverts



High in Openness

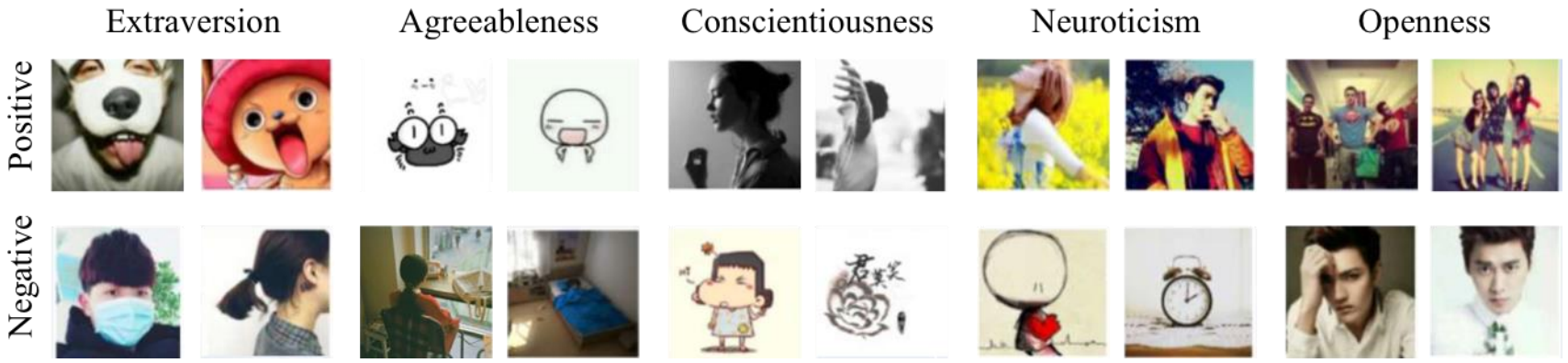
Users high in Openness are more likely to use avatars with their friends.

Users low in Openness prefer avatars with themselves only.

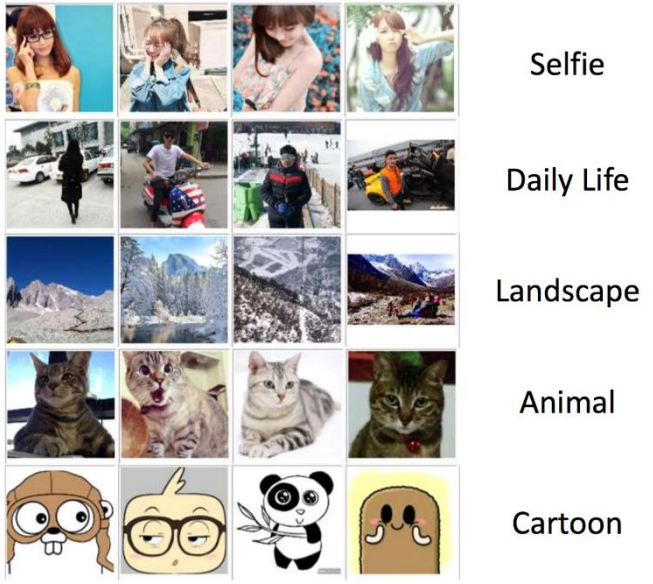
Low in Openness

Avatars

- Avatars matters!



- K-means clustering
  with ResNet results



Emoticons

- Emoticon matters!
  - Agreeable users: smile emoticons
  - Emotionally unstable users: theatrical emoticons to their feelings

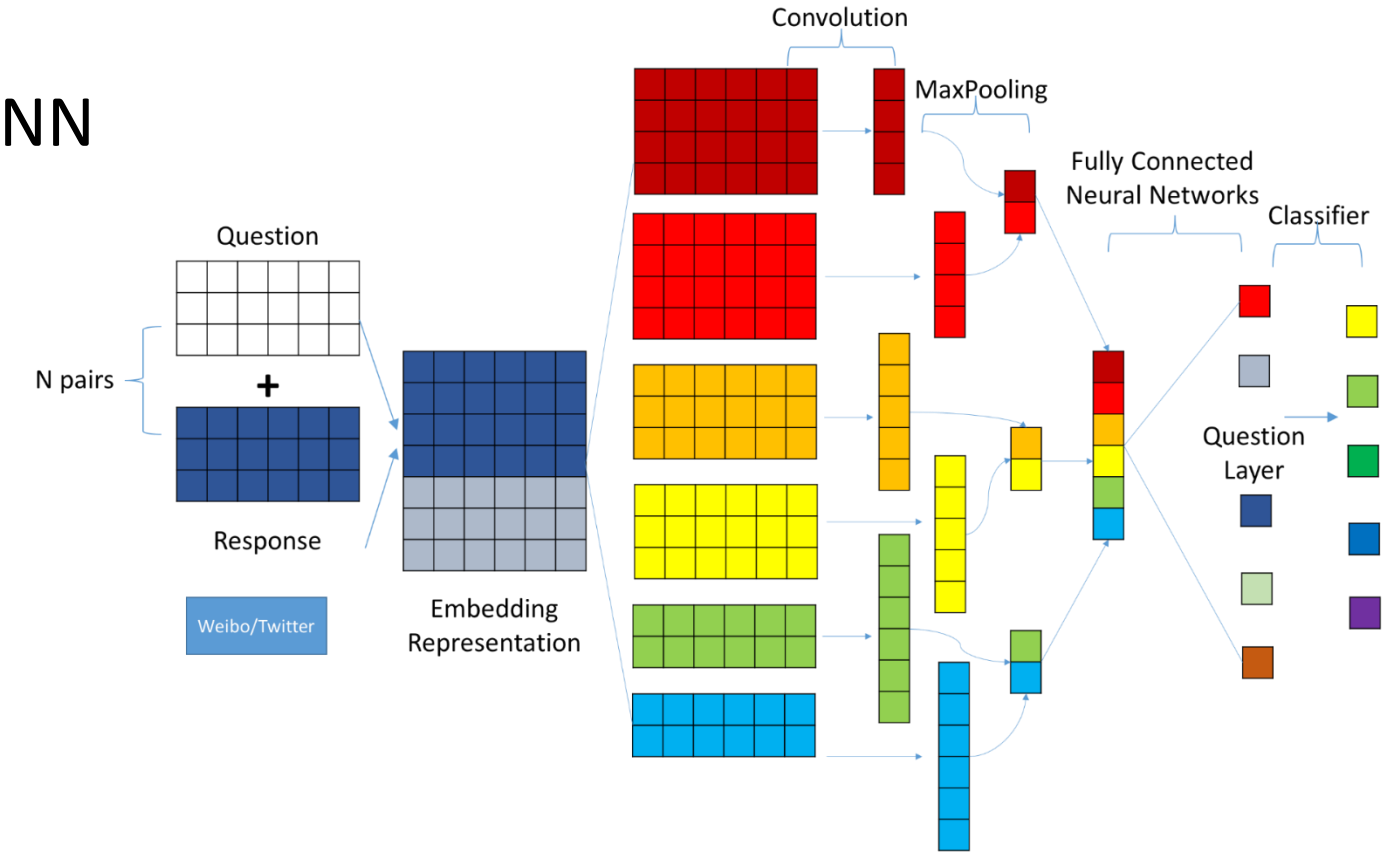
- Pearson Correlation
- Emotion Mapping

Feelings	Emoticon List
Joy	
Anger	
Disgust	
Sadness	
Fear	
Surprise	
Contempt	
Neural	

Responsive Pattern

- Responsive pattern matters!
  - user i tweets a message and his/her fans or followers make
  - user i retweets others' message
  - user i makes comment on others' message

- Responsive-CNN



Evaluation

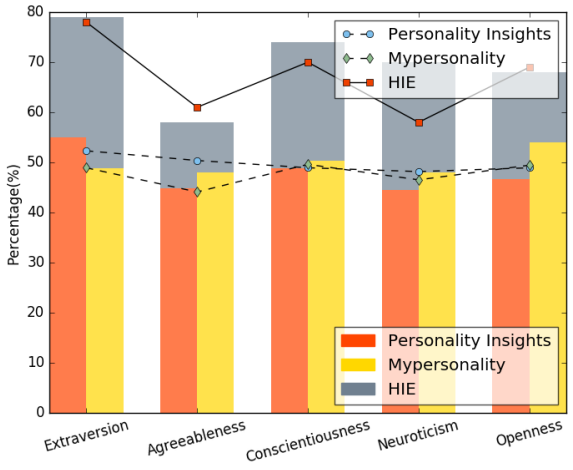
- Personality Segmentations
  - Top or bottom scores in personality traits

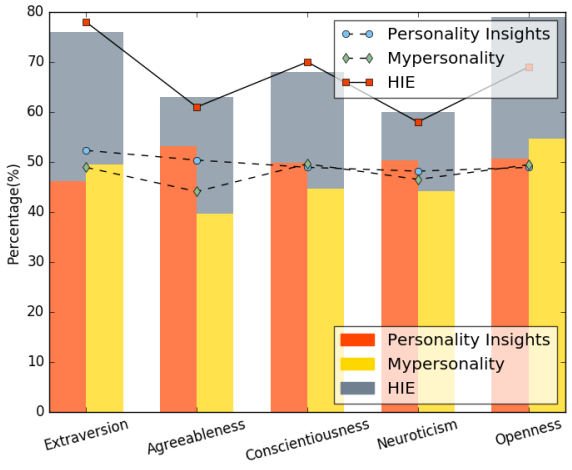
Trait	Positive ($> \bar{x} + \sigma$)	Negative ($< \bar{x} - \sigma$)	Neural
Extraversion	12.4%	13.8%	73.8%
Agreeableness	13.1%	12.6%	74.3%
Conscientiousness	13.4%	12.1%	74.5%
Neuroticism	12.8%	12.8%	74.4%
Openness	12.6%	12.2%	75.2%

- Measure accuracy and precision (as accurate as possible VS as much as possible)
  - Finding the right user is usually far more important than finding as many users as possible in most application scenarios.

Performance and Comparison

- Baseline
  - Text features *(IBM Watson Personality Insights)*
  - Likes and Hashtags *(Facebook Mypersonality)*

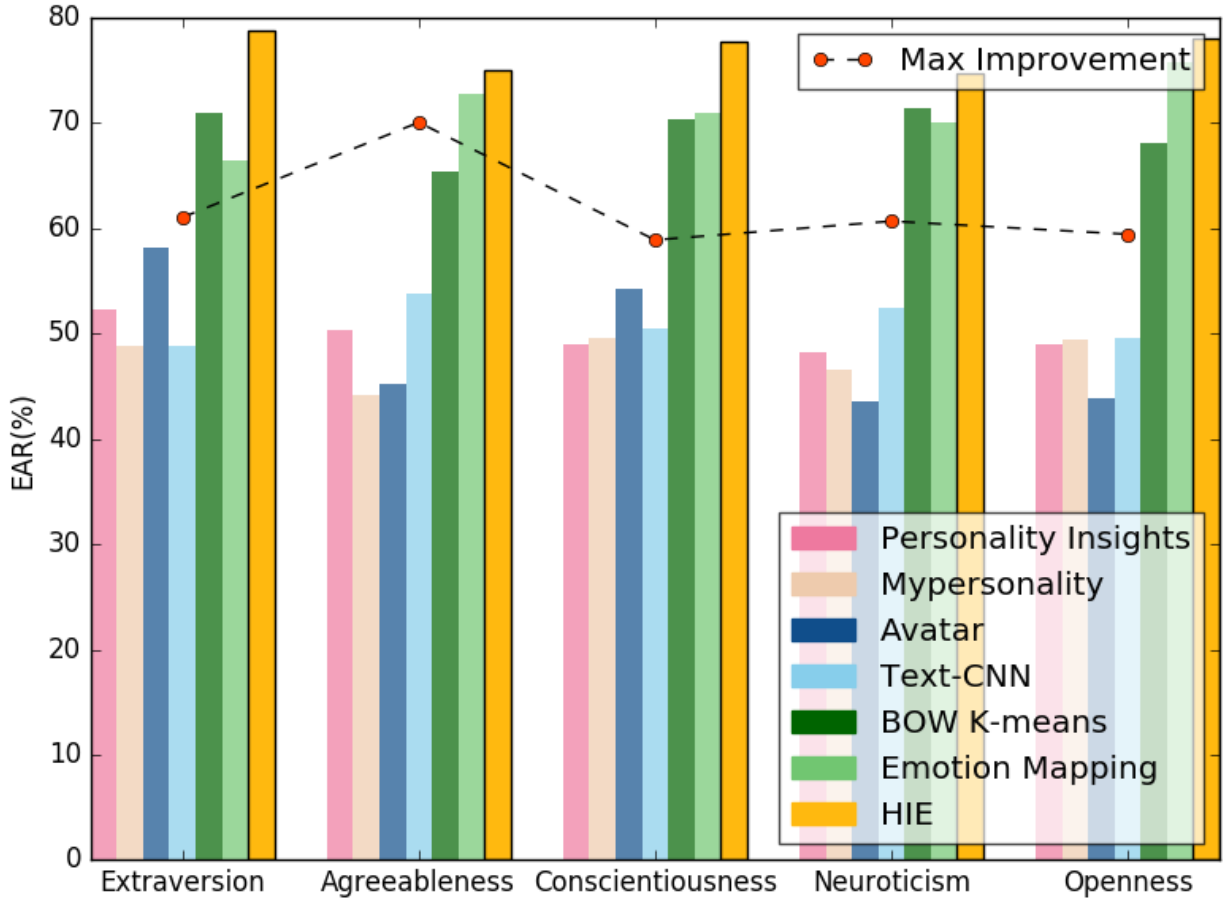




- Raise EPR@P by 61.49% when making predictions on Extraversion.

Advantages of HIE

- Increasing Performance given more features .



Application: Miss Known

Summary

- Propose HIE, a new personality measurement framework
  - Using heterogeneous information in users' digital traces
  - Self-language usage, Avatar, Emoticon and Responsive pattern on social media
- Design various feature engineering strategies
- Applications: MissKnown


Question