Beyond the Words: Predicting User Personality from Heterogeneous Information

Honghao Wei[†,*], Fuzheng Zhang[†], Nicholas Jing Yuan[‡],
Chuan Cao[‡], Hao Fu[‡], Xing Xie[†], Yong Rui[†], Wei-Ying Ma[†]
[†]Microsoft Research [‡]Microsoft
[*]Department of Computer Science and Technology, Tsinghua University
weihh12@mails.tsinghua.edu.cn,
{fuzzhang, nicholas.yuan, chcao, fuha, xingx, yongrui, wyma}@microsoft.com

ABSTRACT

An incisive understanding of user personality is not only essential to many scientific disciplines, but also has a profound business impact on practical applications such as digital marketing, personalized recommendation, mental diagnosis, and human resources management. Previous studies have demonstrated that language usage in social media is effective in personality prediction. However, except for single language features, a less researched direction is how to leverage the heterogeneous information on social media to have a better understanding of user personality. In this paper, we propose a Heterogeneous Information Ensemble framework, called **HIE**, to predict users' personality traits by integrating heterogeneous information including self-language usage, avatar, emoticon, and responsive patterns. In our framework, to improve the performance of personality prediction, we have designed different strategies extracting semantic representations to fully leverage heterogeneous information on social media. We evaluate our methods with extensive experiments based on a real-world data covering both personality survey results and social media usage from thousands of volunteers. The results reveal that our approaches significantly outperform several widely adopted state-of-the-art baseline methods. To figure out the utility of HIE in a real-world interactive setting, we also present DiPsy, a personalized chatbot to predict user personality through heterogeneous information in digital traces and conversation logs.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Data Mining; J.4 [**Social and Behavioral Sciences**]: Psychology

Keywords

User Personality; Big Five; Heterogeneous Information

1. INTRODUCTION

It has been shown in the studies of psychology and behavioral economics that personality, a set of individual patterns and differences based on values, attitudes, personal memories, social relationships, habits, and skills [16], has an effect on individual behaviors, such as occupational proficiency [1] and economic decision [8]. Currently, to measure personality, state-of-the-art surveys or interview-based approaches in psychology rely heavily on retrospective self-reports and thus are vulnerable to memory, not to mention the well-known experimenter effects. In addition, the time and money costs, as well as data granularity, limit the effectiveness and efficiency of these approaches. For example, the 240-item NEO-PI-R [4] and the 300-item International Personality Item Pool (IPIP) [9], require examinees long time to complete time-consuming questionnaires to obtain results. Complicated procedures in data collection hinder the utilization of personality results, such as personalized recommendation services and real-time monitoring for mental health based on psychological features. Therefore, a convincing and effective method of personality evaluation has significant values to psychological studies and relevant Internet services.

In contrast to traditional questionnaire-based methods, psycholinguists suggest analyzing the relationship between users' language features and psychological traits [22, 20]. By understanding the role of language patterns in predicting user personality, computer-based methods are proposed to take users' language usage in social media into consideration, such as the IBM Waston Personality Insights project [10]. With the rise of web and social media, individuals are generating considerable digital traces besides users' language features. Currently, the sequential text data in users' social media contents are utilized for predicting user preferences and characteristics [24, 25]. However, to the best of our knowledge, there has been little exploration of how to predict user personality by leveraging the heterogeneous information embedded in these digital traces.

In this work, we intend to address the following three questions:

The first one is whether self-language usage is the only effective feature to learn user personality. In prior studies, it is mainstream to study users' personality traits by their texts in tweets. LIWC, which is a psychological lexicon, has been used to evaluate user personality according to the usage of words in different semantic categories. Fewer studies have considered other features, such as user likes and hash-tags [23, 19]. However, by observing the social media usage and Big Five[1] personality [12] data collected from thousands of volunteers, we notice curious correlation between heterogeneous information and user personality. For instances, we discover that users with a high score in Agreeableness are likely to use smile

WSDM 2017, February 06-10, 2017, Cambridge, United Kingdom

DOI: http://dx.doi.org/10.1145/3018661.3018717

[1]The five factor model in personality involves dimensions defined as Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness.

emoticons. In contrast, users with a high score in Neuroticism are likely to use theatrical emoticons to exaggerate their feelings. More examples are found, including the relationship between personality and users' gender, place of birth, major, choice of avatars and etc. Similar research has been done in the field of psychology. Taking avatars as an example, it is believed that neurotic participants in games will select avatars with more discrepancies from themselves [3]. Moreover, people with a high score in Openness to new experiences are more likely to choose avatars with fewer discrepancies from themselves [6]. Based on such research, and similar findings in our dataset, we pose a hypothesis that heterogeneous information of users' social media should be leveraged for a better understanding of user personality.

The second one is how to fully leverage the heterogeneous information in users' digital traces. In this work, we present a Heterogeneous Information Ensemble framework, called HIE, to solve this problem. By integrating heterogeneous information including self-language usage, avatar, emoticon, and responsive pattern, HIE first fully leverage the semantic representations learned from the heterogeneous feature engineering part and then apply an ensemble method to integrate all the learned knowledge. We compared the proposed HIE with state-of-the-art computational models from IBM Waston [10] and Mypersonality [23, 19] projects. The results are promising: our approach outperformed the previous methods by up to 70.07% when making prediction on Agreeableness.

Last but not the least, we need to answer the third question: how to apply our approach in a real-world setting. We developed a personalized chatbot, DiPsy, to figure out the utility of HIE in a real-world interactive setting, in which the chatbot reports users' scores in Big Five personality by leveraging heterogeneous information in digital traces, and improves the accuracy of prediction results based on its understanding of interaction through natural conversation logs. Given the unique role of personality in mental health, we plan to build DiPsy as a personalized digital psychologist, who is able to evaluate, diagnose, and treat users' mental process through natural conversations and social media data in the near feature.

In summary, the contributions of this work are as follows,

- We propose a new personality prediction and evaluation framework using heterogeneous information in users' digital traces, including self-language usage, avatar, emoticon, and responsive pattern on social media. To our knowledge, this is the first work to integrate heterogeneous information to study users' personality traits.
- We design heterogeneous feature engineering strategies to extract semantic representations from heterogeneous information on users' social media including tweets, avatars, emoticons and responsive patterns.
- We conduct a large-scale offline evaluation and a real-world service deployment, DiPsy, to study the efficiency and effectiveness of our proposed approaches. The results suggest promising benefits.

2. PROBLEM DEFINITION

In this section, we formalize some concepts frequently used in this work: social media data, personality segmentation, and personality prediction evaluation.

2.1 Social Media Data

Given a user i, we denote the heterogeneous information of this user collected from social media as a set of digital trace instances. It is shown as follows:

$$\begin{aligned} &U_i = \{T_{i,m}, E_{i,n}, A_{i,o}, S_{i,p}\}, \\ &\text{where} \quad m+n+o+p = M, \end{aligned} \tag{1}$$

where $T_{i,m}, E_{i,n}, A_{i,o}, S_{i,p}$ represents an instance of a tweet, emoticon, avatar, and responsive pattern in user i's digital traces respectively. To be specific, responsive pattern refers to how this user responses to other users in social context.

2.2 Personality Segmentation

In the real world settings such as personalized recommendations, the service providers focus on those individuals who are most likely to have interests in their services. Therefore, we focus on users, whose behavior patterns are most likely to be influenced by their personality traits. In Big Five personality measurements, the deviation from the average in scores represents user level of psychological traits and potential impacts of personality in everyday life. Hence, we need to find out those users with top or bottom scores in personality traits, who we denote as positive or negative users, and distinguish them from the rest.

As shown in Table 1, we adopt a three-class segmentation to classify positive, neural, and negative for each dimension in Big Five.

Table 1: Percentage of three-class segmentation in each dimension of Big Five personality.

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%

2.3 Personality Prediction Evaluation

Previous studies in personality prediction mainly evaluate the performance as the correlation between predicted score and ground truth [15]. However, the real world service providers are more interested in the positive or negative users than the overall performance of all users. Taking insurance companies as an example, it is easier to persuade users positive in agreeableness to buy their insurance. By comparison, users negative in agreeableness are less likely to accept the promotion. On top of this, in this work, we focus on distinguishing positive users from negative users, and ignore neural users at most cases. Therefore, our evaluation metric is the binary classification performance between positive user and negative user. First, we use the term "extreme user" to include both positive and negative user, and then define the tailored accuracy and precision metric as follows:

- **Extreme Accuracy Rate** (EAR): The accuracy rate for extreme users is defined as below:

$$\text{EAR} = \frac{\#\text{right predicted extreme users}}{\#\text{total extreme users}} \tag{2}$$

- **Extreme Precision Rate** (EPR): The precision rate for positive and negative users is defined as below:

$$\begin{aligned} \text{EPR@P} &= \frac{\#\text{right predicted positive users}}{\#\text{total predicted positive users}}, \\ \text{EPR@N} &= \frac{\#\text{right predicted negative users}}{\#\text{total predicted negative users}} \end{aligned} \tag{3}$$

We do not evaluate our models by recall because finding the right user is usually far more important than finding as many users as possible in most application scenarios.

3. SYSTEM OVERVIEW

Figure 1 presents the architecture of our system, which consists of three major components: 1) data collection and analysis, 2) heterogeneous feature engineering, 3) multi-classifier ensemble. In data collection and analysis components, we collect the Big Five personality scores of thousands of users and their relevant digital traces on social media. We filter the data and analyze the potential influential factors in personality evaluation, including their usage of language, avatars, emoticons, as well as their pattern of interaction and response in social networks. After that, we use different methods targeted at diverse heterogeneous information of their social media contents in feature engineering layer after the preprocessing, such as Responsive-CNN in understanding the responsive pattern among targeted users and their friends. Last but not least, we leverage stacked generalization-based ensemble method to combine the classifiers in heterogeneous feature engineering layers to compute the final scores of their personality in each dimension respectively. We will detail these components in the following sections respectively.

4. DATA DESCRIPTION

In this section, we present the procedure of data collection and preprocess, and give a brief data analysis.

4.1 Data Collection

We enroll 3,162 users from a medical school in Anhui Province in China. These users agreed to complete a questionnaire evaluating their personalities and provide their corresponding Weibo ID. We take a 44-item Big Five Personality Inventory revised by Oliver [13, 12, 2].

In out data, female users are the majority, comprising 64.78%. Age spans from 17 years old to 25 years old. The average age is 20.84 and the majority of ages are from 20 to 22. Most of the users specialize in nursing (n = 524), clinical medicine (n = 365) and pharmaceutics (n = 342). The Han Chinese is the most dominant ethnicity and the number of Hui people is highest in terms of Chinese ethnic minorities. Most of them come from Anhui, Zhejiang, and Jiangsu Province.

With their Weibo ID and authorization, we crawl users' digital traces for personality prediction. For each of these users, we crawled: a) top 100 tweets recently published on Weibo, b) avatars, c) list of emoticons, d) public profiles, and e) social contexts, including the list of their fans and followers, as well as the tweets with their interaction and involvement.

4.2 Data Preprocess

The credibility of the personality survey results needs to be verified with traits theory and psychometrics. Examined with item-total correlations and factor analysis, the survey results prove simple-structured. We also exempt volunteers whose survey results remain fallacious, such as marking all items with the same scores or reporting similar items with dramatically different answers.

We make a dictionary of 58 words with a prior knowledge and use a lexicon-based method to exclude unqualified texts, such as commercials. In addition, we exempt users who have less than 30 original postings, comments and retweets with their own words. Finally, we enroll 1804 out of 3162 users as volunteers in our study.

4.3 Data Analysis

We analyze the correlation of different aspects from the heterogeneous information on social and user personality. Figure 3(a) shows the top 200 correlated words in Conscientiousness, with the top 10 positive in red and the top 10 negative in blue. We find the words positively associated to Conscientiousness are usually formal words in Journalese, such as "era" and "society", while the negatively associated words are usually informal, such as single characters or typical cyberwords. As shown in Figure 3(b), there are some interesting findings in the usage of avatars and emoticons. For instance, introverts tend to cover their face or show side face. Users high in Openness are more likely to use avatars with their friends, while users low in Openness prefer avatars with themselves only. In addition, users with a high score in Agreeableness are likely to use emoticons with smiles. In contrast, users with a high score in Neuroticism prefer to use theatrical emoticons to exaggerate their feelings.

5. HETEROGENEOUS FEATURE ENGINEERING

In this section, we discuss how to apply different strategies to extract semantic representations from tweets, avatars, emoticons and responsive pattern respectively.

5.1 Tweets

Users' language usage in tweets is the most important indicator of user personality in prior works. Besides the state-of-the-art lexicon-based methods such as LIWC, we also apply three other strategies to analyze the texts, including Pearson correlation, bag-of-words clustering, and Text-CNN as follows:

5.1.1 Pearson Correlation

We use Pearson Correlation to select words strongly correlated with user personality and remove the remaining noisy words. For each dimension in Big Five, we compute the Pearson correlation between each word and the targeted personality, and then select the top 2,000 words.

5.1.2 Bag-of-Words Clustering

LIWC has limited capability to represent users' linguistic patterns in short and informal texts, such as Weibo tweets. Instead we use bag-of-words representation. To be specific, we divide words into five categories including digits, English words, Chinese words, English punctuations, and Chinese punctuations. Then we eliminate all stop words to find out the uniqueness of users' patterns. The results suggest that digits and English words are of low correlation with user personality. Thus, we keep the top 1,500 Chinese words and all the punctuations in the bag-of-words format.

Next, to further alleviate the sparsity problem, we use k-means algorithm to cluster the bag-of-words formated data. After getting the clustering results, we count the number of items within each cluster as the representation of this user.

5.1.3 Text-CNN

Except for hand-crafted features, deep learning has been proved to successfully learn language representation during various NLP tasks in recent years. In our study, we adopt a convolutional network structure to learn tweet representation due to its capability to model the sequential dependency of tweet sentence.

However, standard CNN structure has limited capability to text understanding since there is no direct correlation between the adjacent dimension in embedded word vectors. To solve this prob-

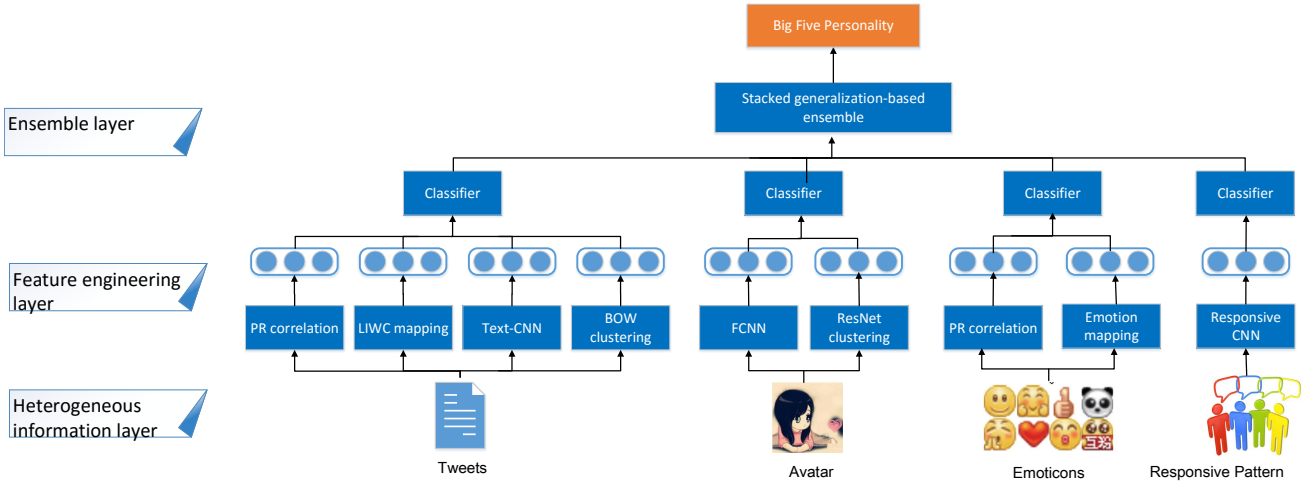
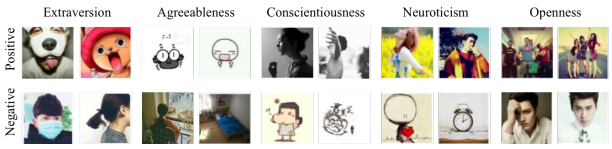


Figure 1: Framework of Heterogeneous Information Ensemble (HIE)

Figure 2: Samples of top correlated items in texts, emoticons and avatars.

(a) Top positively and negatively correlated words in Conscientiousness



(b) Top correlated avatars

(c) Top correlated emoticons

lem, we adopt a graphic model [14] to apply convolution and max-pooling operation with different kernel sizes. All pooling kernels share the same dimension with the input matrix. Therefore, we discard the usage of convolutional layers and concatenate the output of max-pooling to be the input of the next phase in neural networks.

Different from standard CNN, we define feature $c_{i,j}$ to be the i^{th} feature learned from part of the word vector with the j^{th} convolutional kernel.

$$c_{i,j} = f(\sum_{i=1}^{n} W_i x_i + b), n <= 100. \quad (4)$$

where n refers to be the dimension of j^{th} convolutional kernel. For each possible window, we produce a feature map as

$$\boldsymbol{c_j} = (c_{1,j}, c_{2,j}, ..., c_{100-n+1,j}) \quad (5)$$

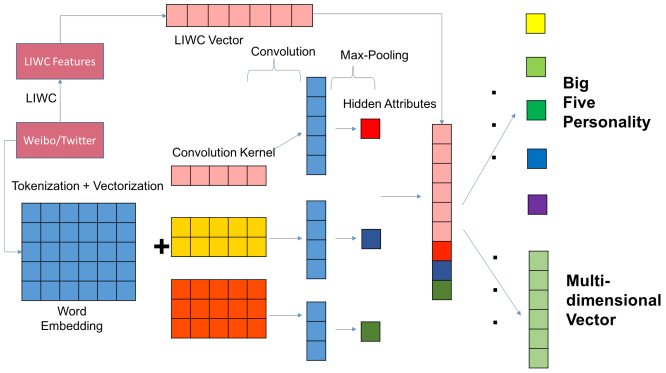


Figure 3: The structure of Text-CNN

Then we apply the max-pooling operation over the feature map and take $\hat{\boldsymbol{c_j}} = max(\boldsymbol{c_j})$ as the pooling results and concatenate them to be the input of next phase $\text{Input}_{\text{next}}$.

$$\text{Input}_{\text{next}} = \hat{\boldsymbol{c_1}} \oplus \hat{\boldsymbol{c_2}} \oplus ...\hat{\boldsymbol{c_N}} \quad (6)$$

where N is the number of convolution and pooling kernel.

The structure of Text-CNN is presented in Figure 3. To better utilize the linguistic information, we incorporate highly correlated LIWC results to the feature vectors before applying softmax activation. Notice the softmax activation is applied to a multi-dimensional output learned with a prior knowledge.

5.2 Avatars

There is little prior work using users' avatars or images to predict personality traits. In section 4.3, we illustrate the potential associations between avatars and user personality.

5.2.1 Fully-Connected Networks

We apply deep learning to learn the nonlinear features in the images, and leverage fully connected neural networks to tackle the embedding results given by the Residential Networking(ResNet) [11]. It assumes the following generative process for each avatar:

1. Resize users' avatars into 256 × 256 pixel matrices.

2. Embed the avatar matrices into 256-dimension vectors by ResNet.

3. Output a non-linear form of hypotheses $h_{W,b}(x)$ with parameters W, b with fully connected neural networks. The hypothesis is defined as $h_{W,b}(x) = f(W^T x) = f(\sum_{i=1}^{256} W_i x_i + b)$ where f is a non-linear function such as the hyperbolic tangent

To better understand user personality, we present the outputs in terms of multi-dimensional vectors with a prior knowledge in psychology. From the personality item pool, we use psychometric theory to select the top 2 most correlated item for each Big Five personality shown dimension as shown in Table 5.2.1:

Table 2: In line with theory in psychometrics, we select the top 2 most correlated items for each dimension in Big Five personality.

Big Five Dimension	Item
Extraversion	Talkative
	Outgoing
Agreeableness	Starts quarrels
	Has a forgiving nature
Conscientiousness	Tends to be lazy
	Does a thorough job
Neuroticism	Relaxed
	Worries a lot
Openness	Imaginative
	Inventive

5.2.2 Clustering in Avatars

We adopt k-means clustering to extract key features in users' avatars as follows:

1. Embed user i's avatar into a 256-dimension vector with ResNet.

2. Given a dataset consisting of 2,000,000 WeChat users' avatars, we adopt the same ResNet for word embedding.

3. Use k-means algorithm to group WeChat users' avatars into 2,000 clusters.

4. Select the closest cluster to representing user i.

5.3 Emoticons

Emoticons are important signals of users' emotions, and thus make an attempt to predict user personality traits via their usage of Emoticons.

5.3.1 Pearson Correlation

This strategy is similar to the one in tweets component. We keep top 50 emoticons strongly correlated with user personality and discard the others.

5.3.2 Emotion Mapping

In our dataset, there are 495 types of emoticons. To enhance the correlation between emoticons and personality traits, we map them into a 8-dimension vector, which represents 8 different emotions respectively. These emotion categories derive from Ekman's atlas of emotions [7], including anger, disgust, fear, joy, sadness, surprise, contempt and neural.

5.4 Responsive Pattern

We emphasize great importance on the responsive pattern. In psychology studies, it is believed that different reactions to the same scenarios during interaction reflect differences in user personality. In this subsection, we introduce Responsive-CNN to learn the semantic representations how user responses to others.

Though the graphic model solves the problem of self-expressed tweets, it ignores the interaction pattern in a social context. Thus, we propose Responsive-CNN, which can capture the interaction pattern between targeted users and their fans or followers on their social network.

We notice that the convolution, pooling and concatenation are operated between different row vectors of the input matrix. In the Text-CNN model, each row vector represents a single word. However, if we embed each tweet into a vector instead of a word, the operation between vectors leads to the interaction of different tweets. We define $q_{i,j}$ as the j^{th} tweet of user i and $r_{i,j}$ as the responding tweets of his/her fans or followers. Then we use $\boldsymbol{qr_{ij}}$ to represent the interaction of user i in following scenarios:

1. user i tweets a message and his/her fans or followers make comment on it.

2. user i retweets others' message.

3. user i makes comment on others' message.

Several methods are applied for tweet embedding, such as using recursive neural networks to transform each tweet into a single vector by word order. However, simple average over word vectors outperforms the others. Therefore, we denote a qr pair as

$$\begin{aligned}\boldsymbol{qr_{i,j}} = \text{average}\{q_{i,j} = (qx_1, qx_2, , qx_{100})\}\cdot \\ \text{average}\{r_{i,j} = (rx_1, rx_2, , rx_{100})\}\end{aligned} \tag{7}$$

We use all qr pairs $\boldsymbol{qr_{i,1}} \oplus \boldsymbol{qr_{i,1}} \oplus \cdots \oplus \boldsymbol{qr_{i,n}}$, where n represents the number of user i's qr pairs, to form the lower part of all user i's matrices. The upper part remains to be word embedding results. We combine the pre-processing and Text-CNN and entitle the new model as Responsive-CNN shown in Figure 5, which use convolution, pooling, and concatenation to operate the interaction between users. In psychological studies, it is explained that the diversity of users' personality traits are revealed from their different reaction to the same content. Therefore, we emphasize importance on the interaction, especially different response to the same contents, to learn differences in user personality.

6. HETEROGENEOUS ENSEMBLE

To fully leverage the semantic representations learned from the heterogeneous feature engineering part. We apply a two-step strategies to predict the user personality.

First, for each domain (Tweets/Avatars/Emoticons/Responsive Pattern), we concatenate the learned representation there to form an integrated vector, and then apply a basic classifier (Logistic Regression) to give the classification score for this domain.

Next, we use stacked generalization-based ensemble method [21], which use the classification score of previous step as input, to give the final result for user personality. This kind of ensemble method can fully leverage the learned knowledge from different aspects and thus archive a better performance for heterogeneous information.

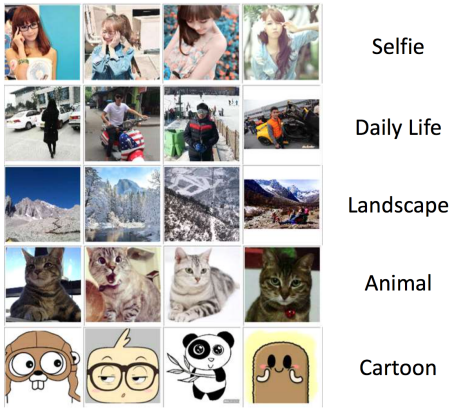
7. EXPERIMENT

We conduct a large-scale offline experiment to study the performance of our approaches for user personality prediction.

Figure 4: Emotion mapping and Avatar clustering.

Feelings	Emoticon List
Joy	
Anger	
Disgust	
Sadness	
Fear	
Surprise	
Contempt	
Neural	

(a) Emotion mapping for emoticons



(b) Sample of clustering for avatars

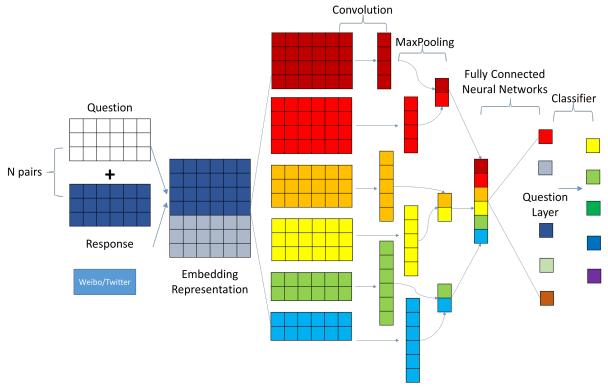


Figure 5: The structure of Responsive-CNN

In this section, we introduce the experiment setting, baseline and ground truth, as well as results of the experiments.

7.1 Settings

We randomly select 70% users to create training set and denote the rest as testing set. For each Big Five personality dimension, we divide the training/testing set correspondingly to ensure the objectiveness. While training, we split another 10% from training set for validation.

7.2 Baseline and Ground Truth

We compare the **HIE** with two baselines, which are the state-of-the-art computational models for user personality using digital traces:

- **Personality Insights**: Personality Insights is deployed on the IBM Waston Developer Cloud to predict user traits in Big Five Personality with users' usage of language. It leverages 256 volunteers' 200 most recent tweets on Twitter, represents the features with LIWC and trains the model with machine learning classifiers. We implement the Personality Insights model on our real-world dataset and train the model respectively with different machine learning classifiers, involving SVM, Logistic Regression, K-Nearest Neighbors, Random Forests, Naive Bayesian and Decision Tree models. The model with logistic regression outperforms the others, and we take it as one of our baseline.

- **Mypersonality**: Mypersonality has been deployed on a Facebook application since 2007 which allows users to test their personality with their Facebook digital traces. It collects personality survey results for 54373 users on Facebook and reports the results with 20 items in IPIP. It leverages the likes mechanism on Facebook and present users' interests with a sparse user-like matrix. Singular-value decomposition (SVD) is used to reduce dimension and the model is trained with Linear or Logistic Regression. We collects Weibo data with users' like information and relevant hash-tags to re-implement it with Logistic Regression.

7.3 Results

In this subsection, we present the result of the **HIE** compared to two state-of-the-art models. We then analyze advantages of the **HIE** and effectiveness of different feature engineering methods.

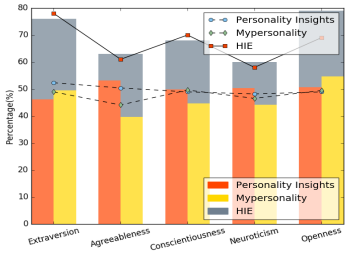
7.3.1 Overall Results

Figure 6 shows the performance of the HIE (with tweets and avatars only) and the two aforementioned baselines, where HIE outperforms other methods in terms of EAR, EPR@P, and EPR@N significantly. To be specific, HIE outperforms the previous methods by up to 61.49% in EPR@P when making predictions on Extraversion. The results demonstrate the advantages of the HIE.

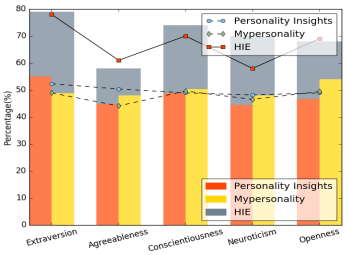
Noticing the performance of LIWC and user-like matrix methods, we find the aforementioned approaches problematic in our data. Generally speaking, the optimum setting for LIWC is long and formal texts in line with grammatical and lexical rules. However, referred to as representation of short texts and imformal words, tweets fail to comply with the strict rules in linguistics. We calculate the Pearson correlation of the LIWC items and user personality. Linguistic rules in LIWC are not presented well in results. There is much lower correlation compared to items in bag-of-words and emotion mapping. Ironically, the top correlated items in LIWC are observed in contradiction with the rule itself. For instance, death words are one of the most positively correlated items to Agreeableness and the exclamation mark is negatively related to Openness.

With heterogeneous information in users' digital traces, **HIE** successfully solves these problems. We take features beyond the usage of language to better understand user personality and collect

Figure 6: EAR, EPR@N and EPR@P of HIE (with tweets and avatars) and the two aforementioned state-of-the-art models.



(a) EAR and EPR@N



(b) EAR and EPR@P

information in different aspects. In the next part, we would explain the advantages of the **HIE**.

7.3.2 *Advantage of HIE*

To further study the advantage of the **HIE**, we test the classifiers with the same input in terms of bag-of-words and emoticons. As shown in the Figure 8(a), by adding the HIE into the classifiers to better represent users' digital traces, we notice a significant improvement over EAR, EPR@P, and EPR@N. This validates the ability of HIE to predict user personality from heterogeneous information.

Due to the demands for personalized services, reliability of predicted results is more important than detecting all potential individuals in a specific personality category. **HIE** perfectly meets this need of real-world usage. Depicted in Table 7.3.3, we notice that users with high probability scores are aggregated in the two ends of the confidence interval. Moreover, users with extremely low scores are on the left end while users with extremely high scores are on the right end. It shows the ability of HIE to accurately distinguish the targeted extreme users from the others. In addition, Figure 8(b) shows the distribution of the rest of users with medium scores in personality. It reveals that **HIE** is capable of aggregating users' with less noticeable personality features in the middle intervals.

There is a question about modules under the **HIE**: with the integration of more modules from heterogeneous information, would the performance of HIE increase? To answer the question, we compare the HIE using single module with the one using an ensemble of diverse features, including tweets (BOW clustering), avatars (ResNet and fully connected neural network), emoticons (emotion mapping), and responsive pattern (Responsive-CNN). Presented in the Figure 8, there is a noticeable increase in performances which bolsters the potential of our framework.

7.3.3 *Results for Feature Engineering*

Due to the lack of space, we only present the results for Text-CNN and Responsive-CNN. We notice a significant margin between Text-CNN and other text-processing methods.

We also find a significant increase in performance of Responsive-CNN compared to Text-CNN, which leverages the similar CNN structures, presented in the Figure 9. We believe the responsive pattern and interaction with friends are as important as language itself in inferring user personality, especially when the targeted user responds differently to the same items. Figure 10 presents an example, where the user posts a tweet with a smiling emoticon and lovely selfies to show her agreeableness. However, in replying to the comments of the posting, she behaviors as an angry woman with unstable emotional status and use an angry emoticon to reveal her feelings directly. By integrating the contradictory performances during interaction, Reponsive-CNN is likely to have an overall view on user personality.

Table 3: Distribution of Positive, Negative and Neural users in confidence intervals (presented in terms of percentage).

Interval	[0,0.1]	[0.1,0.2]	...	[0.8,0.9]	[0.9,1.0]
Extraversion					
Negative user[1]	97.26	88.89	...	28.57	2.99
Positive user[2]	2.74	11.11	...	71.43	97.01
Neural user[3]	5.47	7.76	...	2.38	1.82
Agreeableness					
Negative user	98.36	77.14	...	19.23	7.69
Positive user	1.64	22.86	...	80.77	92.31
Neural user	4.41	4.91	...	6.13	6.56
Conscientiousness					
Negative user	100.00	93.75	...	22.22	4.29
Positive user	0.00	6.25	...	77.78	95.71
Neural user	2.38	1.84	...	8.35	19.67
Neuroticism					
Negative user	93.42	80.00	...	20.00	1.54
Positive user	6.58	20.00	...	80.00	98.46
Neural user	17.05	10.80	...	1.18	0.75
Openness					
Negative user	97.30	79.31	...	11.43	2.94
Positive user	2.70	20.69	...	88.57	97.06
Neural user	12.55	11.44	...	1.44	0.92

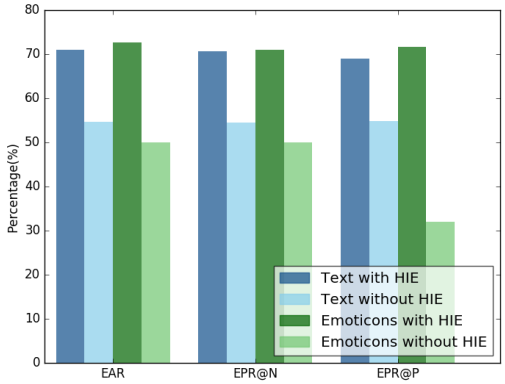
8. DIPSY, A CHATBOT FOR PERSONALITY PREDICTION

To test the reliability of our framework in a real-world setting, we design and build a chatbot, called DiPsy to evaluate user personality through digital traces and conversation logs. The system consists of four phases: Onboarding, Conversation, Reporting, and Diagnosis & Treatment.

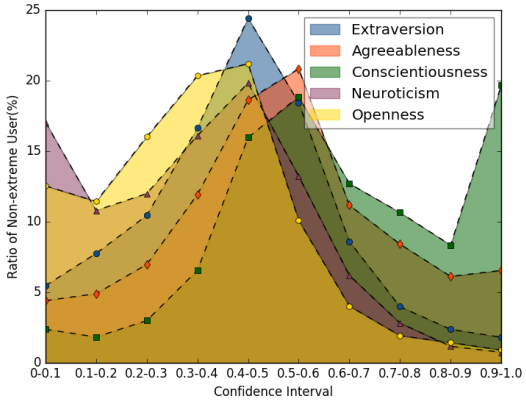
8.1 Onboarding

During the onboarding process, each user opts-in their Weibo and WeChat accounts to allow access to their digital traces. These traces will be securely pulled from services with appropriate authorization enabled by OAuth protocol provided by Sina Weibo. We extract users' heterogeneous information on social media, including their language usage, avatars, emoticons, individual demographics, friendship on social media, interaction, and responsive patterns. Based on our approaches, we evaluate user personality and create a profile for each individual user with their personality traits analyzed from digital traces.

Figure 7: EAR, EPR@N and EPR@P of logistic regressions with and without the HIE in terms of tweets and emoticons and Distribution of neural users in confidence intervals.



(a) Tweets and Emoticons with & without HIE.



(b) Distribution of neural users in confidence intervals.

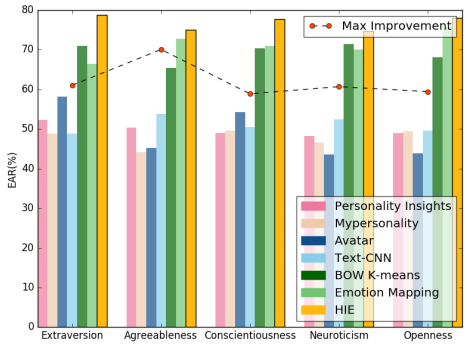


Figure 8: Comparison about performance of different models over EAR.

Figure 9: EAR of Resposive-CNN and Text-CNN in 10 IPIP items

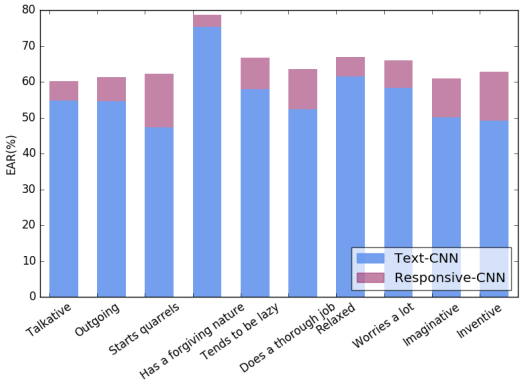
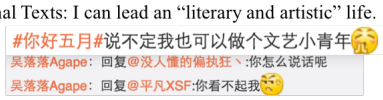


Figure 10: Example of responsive-patterns

Original Texts: I can lead an "literary and artistic" life.



Replies: How dare you say to me like that!
You despise me (with an angry emoticon)!

8.2 Conversation

After HIE analyzes user personality based on heterogeneous information in digital traces, DiPsy presents tips for usage to encourage users for a conversation. In fact, we integrate the natural conversation logs to revise the results of user personality. In this phase, DiPsy runs as a typical chatbot. We have a background chat engine to conduct chit-chat with users. Given users' input, DiPsy will decide whether the input can be used to perform psychological diagnosis and help generate the response for guiding new conversations. If the input text is not relevant for psychological diagnosis, a chit-chat response will be generated using the background chat engine. Otherwise, conversation information, which represents the interactive process between DiPsy and users, is translated into the characteristics in the feature layer with deep learning techniques, such as Responsive-CNN.

There are several other strategies for determining user personality. First, we would introduce 'chicken soup for the soul' to the users and observe their reactions. By mapping their answers to highly correlated items, we can link text information of users' tweets to their attributes in Big Five personality. Second, when results are presented with high confidence, DiPsy would explicitly mentions highly correlated items from the personality item pools. The items are presented in the form of words and audio, such as the audio to ease nerves. The positive or negative answers to the item directly help to revise the observation over user personality. Last but not the least, we specifically detect key words through conversation. For instance, if word ḋepressionánd áutismáre detected from the answer, the targeted users are more likely to have high scores in Neuroticism.

8.3 Reporting

After revisal through conversation, we present user personality when the confidence of analysis exceeds certain thresholds or the user chooses to opt out of DiPsy. The results are shown in both texts and images. We use two radar graphs to demonstrate the different

scores in each Big Five personality dimension and top 10 correlated items. We also provide explanations to help users interpret the results, together with advice for daily-life based on personality profiling. For instance, we encourage introverts to make more friends and participate in outdoor activities outside of work and school.

8.4 Diagnosis and treatment

Currently, DiPsy serves as a personalized chatbot to evaluate user personality. Due to the wide use of personality in mental health examination, we hope to build DiPsy as a digital psychologist to diagnose, and treat users' mental process through digital footprints and natural conversations in the near future. By determining users' mental status, DiPsy is able to conduct cognitive behavioral therapy (CBT) or early intervention to help at-risk users alleviate/manage their problems by changing the way they think and behave in a variety of therapeutic contexts.

In this way, this digital chatbot can be used for both consumer and enterprise. DiPsy creates a new paradigm for human resources management, which provides a more natural way to 1) help organizations profile employees based on their psychological traits, 2) guide the team structuring for a better team chemistry and productivity, and 3) coach employees for mental health care on a daily basis.

9. RELATED WORK

Broadly, this study falls into the category of the training models that use users' digital traces to evaluate personality traits. The state-of-the-art computational models involve questionnaire-based [1, 5] and data-driven approaches [15, 23, 19, 10]. The Prior work with data-driven methods has considered information including texts [10], interests [19], hash-tags and likes [15, 23]. However, most prior work in this line of research only considers simple text information and processes it with closed vocabulary, such as Linguistic Inquiry and Word Count(LIWC) [17], and open vocabulary methods, such as topic modelling. To the best of our knowledge, little prior work considers user behavior patterns from the angle of diverse features in heterogeneous information on social media. In our study, we apply different strategies to extract semantic representations from heterogeneous information in user digital traces and represent the diverse patterns embedded in it. For instance, we leverage Responsive-CNN to better understand the interaction among users and K-Means clustering with ResNet [11] embedding results to process avatars.

In our work, we also propose a new heterogeneous information ensemble framework, called HIE. Users' tweets, avatars, emoticons, individual profiling, social media friends, and responsive-patterns, are integrated together to improve the performance. In contrast, prior work only considers limited genres of users' digital footprints [15, 23]. Moreover, those methods targeted at single feature, such as latent Dirichlet allocation (LDA) [19], fail in our case, because the diversity of users' behavior patterns directs the clustering results to be of low correlation with targeted features. We apply a two-step strategy with stacked generalization-based ensemble method to achieve a better performance.

Finally, with the popularity of combining the knowledge of psychology and the technology of computer science, the emergence of mobile psychology and cyber psychology draws people's attention [18]. To figure out the utility of HIE in real world interactive settings, we build DiPsy, a personalized chatbot, which reports users' scores in personality by leveraging heterogeneous information on social media, and revises the results based on its understanding of the interaction during natural conversations. To the best of our knowledge, there is little prior work capable of learning users' psychological traits via digital footprints and revising results with dialog information.

10. CONCLUSION

In this paper, we propose a heterogeneous information ensemble framework, called HIE, to predict user personality by integrating heterogeneous information in digital traces including self-language usage, avatars, emoticons, and responsive patterns. In HIE, different strategies are applied to extract semantic representation for heterogeneous information and stacked generalization-based ensemble method is applied to fully leverage the learned representation. Extensive experiments and analysis have been done on a real-world dataset covering both personality survey results and social media usage from 3,162 volunteers. The results are promising and HIE outperforms the state-of-the-art models in all Big Five personality dimensions. Last but not the least, we developed a WeChat mobile application, DiPsy, to evaluate user personality to demonstrate the reliability of HIE in real-world settings. We hope to develop DiPsy to diagnose and treat users' mental process through natural conversations and social media data in the near future.

11. REFERENCES

[1] M. R. Barrick and M. K. Mount. The big five personality dimensions and job performance: a meta-analysis. *Personnel psychology*, 44(1):1–26, 1991.

[2] V. Benet-Martinez and O. P. John. Los cinco grandes across cultures and ethnic groups: Multitrait-multimethod analyses of the big five in spanish and english. *Journal of personality and social psychology*, 75(3):729, 1998.

[3] K. Bessière, A. F. Seay, and S. Kiesler. The ideal elf: Identity exploration in world of warcraft. *CyberPsychology & Behavior*, 10(4):530–535, 2007.

[4] P. T. Costa and R. R. MacCrae. *Revised NEO personality inventory (NEO PI-R) and NEO five-factor inventory (NEO FFI): Professional manual*. Psychological Assessment Resources, 1992.

[5] P. T. Costa and R. R. McCrae. *Neo Pi-R*. Psychological assessment resources, 1992.

[6] R. A. Dunn and R. E. Guadagno. My avatar and me–gender and personality predictors of avatar-self discrepancy. *Computers in Human Behavior*, 28(1):97–106, 2012.

[7] P. Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200, 1992.

[8] J. K. Ford. *Brands laid bare: Using market research for evidence-based brand management*. John Wiley & Sons, 2005.

[9] L. R. Goldberg, J. A. Johnson, H. W. Eber, R. Hogan, M. C. Ashton, C. R. Cloninger, and H. G. Gough. The international personality item pool and the future of public-domain personality measures. *Journal of Research in personality*, 40(1):84–96, 2006.

[10] L. Gou, M. X. Zhou, and H. Yang. Knowme and shareme: understanding automatically discovered personality traits from social media and user sharing preferences. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, pages 955–964. ACM, 2014.

[11] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.

[12] O. P. John, E. M. Donahue, and R. L. Kentle. The big five inventoryâĂŤversions 4a and 54, 1991.

[13] O. P. John, L. P. Naumann, and C. J. Soto. Paradigm shift to the integrative big five trait taxonomy. *Handbook of personality: Theory and research*, 3:114–158, 2008.

[14] Y. Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.

[15] M. Kosinski, D. Stillwell, and T. Graepel. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110(15):5802–5805, 2013.

[16] W. Mischel, Y. Shoda, R. E. Smith, and F. W. Mischel. Introduction to personality. *University of Phoenix: A John Wiley & Sons, Ltd., Publication*, 2004.

[17] J. W. Pennebaker, M. E. Francis, and R. J. Booth. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71:2001, 2001.

[18] J. Piazza and J. M. Bering. Evolutionary cyber-psychology: Applying an evolutionary framework to internet behavior. *Computers in Human Behavior*, 25(6):1258âĂŞ1269, 2009.

[19] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M. E. Seligman, et al. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791, 2013.

[20] Y. R. Tausczik and J. W. Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54, 2010.

[21] D. H. Wolpert. Stacked generalization. *Neural networks*, 5(2):241–259, 1992.

[22] T. Yarkoni. Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. *Journal of research in personality*, 44(3):363–373, 2010.

[23] W. Youyou, M. Kosinski, and D. Stillwell. Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences*, 112(4):1036–1040, 2015.

[24] Z. Zhao, H. Lu, D. Cai, X. He, and Y. Zhuang. User preference learning for online social recommendation.

[25] Z. Zhao, Q. Yang, D. Cai, X. He, and Y. Zhuang. Expert finding for community-based question answering via ranking metric network learning.