Soc. Netw. Anal. Min. (2016) 6:74
DOI 10.1007/s13278-016-0383-4

ORIGINAL ARTICLE

Identifying and validating personality traits-based homophilies for an egocentric network

Md. Saddam Hossain Mukta[1] · Mohammed Eunus Ali[1] · Jalal Mahmud[2]

Received: 27 May 2016 / Revised: 5 August 2016 / Accepted: 24 August 2016 / Published online: 6 September 2016

Abstract Social network sites (SNS) have touched the lives of millions of people around the world. People share interests, ideas, photos, activities in the social networks with their family, colleagues, friends and acquaintances. However, the degree of interactions among members widely varies. According to a sociology principle, people with similar personality often interact with each other more frequently. A group of connected people with similar personality traits is termed as a *homophily*. In this paper, we develop a method to identify homophilies by analyzing the Big5 personality traits of users from their interactions in an egocentric network like Facebook. We observe that our homophilies correctly cluster ranged from 73 to 87 % users for different personality traits. We also present a novel validation technique to verify those extracted homophilies in real life. Note that we are the first to validate the extracted homophilies and compare those with baseline techniques from SNS usage in real life using an interview-based method. We notice that our validation results show different agreements ranged from 0.207 (fair) to 0.709 (substantial) among the raters of those homophilies in real-life .

✉ Md. Saddam Hossain Mukta
saddam944@gmail.com

Mohammed Eunus Ali
eunus@cse.buet.ac.bd

Jalal Mahmud
jumahmud@us.ibm.com

[1] Department of Computer Science and Engineering, Bangladesh University of Engineering and Technology, Dhaka 1000, Bangladesh

[2] IBM Research-Almaden, San Jose, CA, USA

Keywords Regression · Classification · Clustering · Intra-class correlation

1 Introduction

People are increasingly using social network sites (SNS) such as Facebook and Twitter to share their thoughts with family, friends and acquaintances. Recent research shows that it is possible to capture individual behavioral attributes, e.g., personality, from their interactions in SNS. Personality of an individual is commonly defined using five psychological traits: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism, also known as Big5 John et al. (2008). In this paper, we first identify communities based on the similarities of Big5 personality traits, which we called *homophilies*, from social media (i.e., Facebook) usage. We also devise a questionnaire-/interview-based validation technique, which enables us to verify the identified homophilies in real life. Moreover, we compare our approach with two baseline homophily identification techniques. Note that, we focus on identifying and validating homophilies in an *ego-centric network* Fisher (2005), where the ego-centric network denotes social relationship between an ego (a user) and other connected members (Facebook friends) with the ego in the network.

Homophily is a group of people who are strongly connected with each other. Authors in a sociology study McPherson et al. (2001) described that homophilies can be identified based on different cognitive attributes of humans such as attitude, belief, behavior and so on. In another sociology study Back et al. (2010), authors demonstrated that Facebook reveals actual personality of people. During Facebook interaction, people cannot reveal

different personality traits apart from their own. Authors Kosinski et al. (2013) also successfully made a predictive analysis from Facebook *likes* about different highly sensitive personal attributes including ethnic origin, political views, religion, satisfaction in life and substances use (e.g., alcohol). Facebook *likes* also express users' positive association with online contents such as product, restaurant, sports or music. The above studies motivate us that it is possible to extract actual personality trait-based homophilies from social media usage in real life.

Though a large body of works on identifying the personalities of users from social media usage has been appeared in recent years Schwartz et al. (2013), Golbeck et al. (2011), identifying homophilies in social medias has not been studied until recently. Kafeza et al. (2014) first proposed a method to identify influential communities in Twitter based on the users' personality traits. In an open network like Twitter, message or content is made available to everyone who wants to receive it. Kafeza et al. (2014) identify influential communities in Twitter. Since, they extract influential communities of the entire network, these communities largely represent celebrities, domain experts, and scientists. However, these people may not be friends to each other in real life. Thus, it is very difficult to validate whether personality of members are similar to each other in an open network. Hence, authors' Kafeza et al. (2014) approach limits applicability in the practical world. On the contrary, we identify personality trait-based communities for each member considering him as an ego in an egocentric network. In an egocentric network (e.g., Facebook), a connection is made between participants when they are agreed to do so. Previous trust among participants leads them to be connected in the egocentric network Petrocelli (2014). When an ego is changed, our approach reshapes community according to his personality traits. Since, the network is closed (i.e., Facebook), it is possible to validate similarity of personality among members of the extracted communities in real life. We are the first to validate personality traits derived from social network which conforms with the traits in real life.

Finding the personality trait-based homophily in an egocentric network like Facebook has a number of real-life applications. Let us consider the following scenario. A person sends a recommendation to another person (friend) in the Facebook. If the second person sees that the personality of the recommender matches with her own and the recommender is already known to her (trusted) in real life, then there is a high chance that the second person will accept the recommendation. Other applications that can be benefited from personality traits-based homophilies identification in an egocentric network include finding similar acquaintances in workplaces, observing the social dynamics (e.g., group behavior), parental approval in friends searching, accepting friend requests, etc., Bisgin et al. (2010).

Our proposed approach has two major steps: (1) identifying homophilies from users' Facebook usage, and (2) validating the identified homophilies in real life using questionnaires.

In the first step, we build a homophily prediction model by analyzing the Facebook usage (statuses) of users. In this process, we first analyze statuses of 663 Facebook users by using both closed [e.g., Linguistic Inquiry and Word Count (LIWC)] Gilbert and Karahalios (2009) and open [e.g., Meaning Extraction Helper (MEH)] Schwartz et al. (2013) vocabulary-based approaches. Then, we conduct a standardized 44 items IPIP John (2000) personality test to compute Big5 John et al. (2008) scores of each user, which we consider as the ground truth data of user personality. Later, we build stepwise forward selection-based automatic linear regression model, where a user's Facebook statuses are given as input and Big5 personality scores are produced as output. We observe that LIWC-based approach produces moderate strength while predicting personality score due to its less expressiveness (i.e., capture only 4500 dictionary words and word stems). Later, we build our personality prediction model with two linguistic features (i.e., '1-g' words and topics) using open vocabulary-based approach. Next, we compare the strength of these two models, i.e., open and closed vocabulary-based approaches. We observe that open vocabulary-based approach with two linguistic features (i.e., '1-g' and topics) outperforms closed vocabulary-based approach. Then, we combine these two linguistic feature-based models and use a linear weighted ensemble-based technique to compute final personality score that produces significant improvement over single linguistic feature (i.e., '1-g' words or topics)-based approach Sill et al. (2009) and Jiménez (1998). By using these ensemble-based personality scores, we identify homophily of an ego and all connected users (friends) from their Facebook usage.

In the second step, we develop a questionnaire-based novel validation technique that enables us to validate the identified homophilies in real life. For this step, we need to collect the data of Facebook friends who are also friends in real life. We collect their Facebook usages and construct the homophily network by using our prediction model developed in the first step. Our homophilies correctly cluster 155 Facebook users of validation dataset ranged from 73 to 87 %. After that we conduct individual trait-based IPIP test on these users, where everyone rates himself and other connected friends in terms of personality scores through different questionnaires. To validate homophily result statistically, we compute *intraclass correlation coefficient (ICC)* Koch (1983) and *Cohen's*

Kappa Viera et al. (2005) values among members. We find several *moderate* and *substantial ICC* scores between our computed personality scores and self evaluating result by the homophily members in real world. We compute Cohen's Kappa to measure inter-rater agreement between two observers about homophily members. Later, we compute two baseline homophily identification techniques from the existing studies and make a comparison whether our approach can more accurately determine personality trait-based homophilies than those techniques. We also propose a group recommendation technique. The technique recommends movie preference to a group of users who are similar to an ego. We find strong correlation between trait-based personality scores of homophily members and their movie watching preferences in real world.

In summary, our contributions are as follows:

- We identify multiple personality trait-based homophilies in an egocentric network, where five different personality traits (i.e., Big5) are considered as five different homophilies.
- We compare the strength of the models (i.e., open and closed vocabulary) and select the one that predict better personality scores.
- We first build an ensemble-based personality identification technique with two different linguistic features, i.e., '1-g' words and topics.
- We develop a novel interview-based homophily validation technique to measure the accuracy of our framework.
- We present experimental result to show the accuracy of our homophilies and compare those homophilies with other baseline techniques.
- We propose a group recommendation technique based on Big5 personality trait.

The remainder of this paper is organized as follows. Section 2, discusses the problem formulation, and Sect. 3 describes the related work. In Sect. 4, we present data collection process. Section 5 shows personality building model, and Sect. 6 reports model selection process. Sections 7 and 8 describe homophily identification and validation techniques, respectively. Sections 9 and 10 present baseline and group recommendation techniques, respectively. In Sect. 11, we present discussion of our experiments and findings. Finally, Sect. 12 concludes the paper with a discussion of future research direction.

2 Problem formulation

Let $G = (V, E)$ be an undirected social network where V and E denote the set of vertices (users) and set of edges (social connections), respectively. We consider the social network graph as *Big5* personality graph. Each vertex (user) in G has *Big5* personality attribute set $V.a \in P$, where $P = \{o, c, e, a, n\}$ denotes five different traits of personality and $V.a$ is a real value in the range of [0, 1] . In this section, we first give necessary definitions and then state our problem statement.

Definition 1 *Egocentric social graph* Let u be an ego, and G be the social graph. Then, an egocentric graph G_u of u can be defined as follows: $G_u = (V_u, E_u)$ be an undirected connected egocentric social network where V_u and E_u denote the set of vertices and set of edges, respectively, connected to ego u. Here, $u \in V_u$ be the ego of the network where $V' = V_u$ u denotes the alter of the graph.

Definition 2 *Homophily graph* Homophily is a subgraph $H^{u\delta}$ of G_u for ego u with respect to a personality attribute $\delta \in P$, where the difference of the personality traits of the ego and any other member $u' \in H_u$ is less than ϕ, i.e., $|u.\delta - u' \cdot \delta| < \phi$. We get *five* different homophilies $H_{u\delta}$ of ego u, where $\delta = 1, 2, \ldots 5$.

The goal of this study is to identify personality trait-based homophilies of an egocentric social network. We make a predictive modeling of personality trait-based homophily whether users are similar to an ego. For example, we find five different homophilies $H_{\text{uOpenness}}, H_{\text{uConcenticiouness}}, H_{\text{uExtraversion}}, H_{\text{uAgreeableness}}$, and $H_{\text{uNeuroticism}}$ of an ego u according to Fig. 1. We also validate these homophilies in real world.

3 Preliminaries and related work

3.1 Big5 model

Big5 model is one of the well-studied topics in personality research Norman (1963). Big5 model has five personality traits: Openness, Conscientiousness, Extraversion,

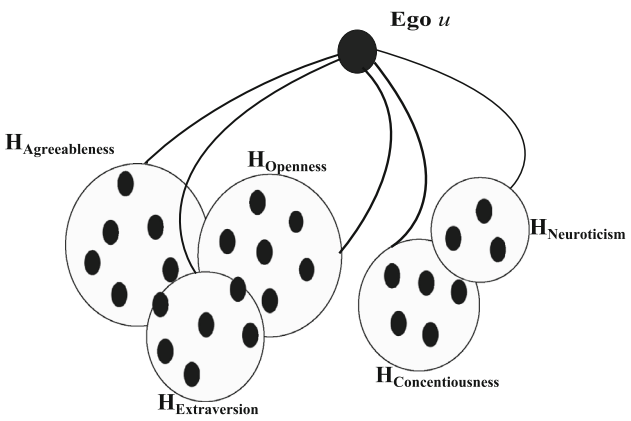

Fig. 1 Personality trait-based multiple homophilies in an egocentric network. Ego u finds different homophilies based on his five different personality traits

Agreeableness, and Neuroticism. Sociology studies show that Big5 traits are consistent across gender, age and languages Golbeck et al. (2011), Yarkoni (2010), John and Srivastava (1999) to analyze human personality. Big5 model is distinguished as follows John and Srivastava (1999), Openness to Experience trait has tendency to reflect ideas, innovation and appreciation for values. People with high Conscientiousness trait are prone to be cautious, meticulous and a tendency to seek achievement. Extraversion trait has tendency to seek excitement and show positive emotions. People of Agreeableness trait tend to be sympathetic, trusted and merciful to others. Neurotic individuals show negative emotions such as anxiety, inhibition, anger, and depression.

3.2 Related work

3.2.1 Homophily and tie strength

We share our intimate personal information in social networks Schwartz et al. (2013). These networks become ideal place for testing our social phenomena (e.g., homophily) McPherson et al. (2001), Crandall et al. (2008). Authors in Bisgin et al. (2010) investigated homophily upon interests of individuals such as genre of music, blog, tags, categories, and friends they like. They studied similarity of members by computing Jaccard similarity co-efficient. Authors Adali and Golbeck (2012) described person's social circle of similar people (*homophily*) from different action features (i.e., network bandwidth, message content, pair behavior, etc.) of a public network like Twitter. Gilbert and Karahalios (2009) defined a predictive model to find out weak and strong tie among the users of a social network. Trusted family and friends are categorized as strong tie. Authors Gilbert and Karahalios (2009) experimented with 2184 friendships for collecting data with a random subset of friends by *Greasemonkey*, an automated script which was executed at client side. They identified 74 Facebook variables such as predictive variables (intensity, intimacy, duration, structural, emotional support and social distance), structural and dependant variables. Finally, the score scale lie 0–1 which is capable of finding the weakness or strengthens of ties. It is possible to find out tie of individual and clique, but they did not focus on extracting tie strength upon personality matching.

3.2.2 Social circles

Valerio et al. (2013) explained relation to human behavior and dynamics of interaction over time. They observed a large percent of weak ties and turnover among users in Twitter can be modeled how society is changing. According to anthropology and psychology literature, they modeled different types of egocentric networks, such as support clique (5 members), sympathy (15 members) group, affinity (50 members) group and active network (150 members). Authors found different psychological circles which may evolved over time. They also discovered similarity among the alters and ego from different perspectives (i.e., sphere of sentiment, intensity of activities). However, the authors did not consider similarity among members inside the circles in terms of personality. Authors in Crandall et al. (2008) described that people are similar to their neighbor in a social network for two different reasons: (1) *social influence* guide them to adopt similar behavior of their neighbors and (2) people likely to establish new relationship who are already similar to each other. Authors use wikipedia dataset to find out homophily of users based on users' activities (i.e., editing an article) and interactions (i.e., talk to a person). McAuley and Leskovec (2012) described that SNS are growing enormously and it is cumbersome to customize member list (e.g., Google+ circle, Facebook friend list) manually. Authors proposed an automated circle exploring approach by discovering common aspect (e.g., family members, university friends) of a group of people. These circles could be nested, hierarchical and overlapping of members. But authors did not extract any circle based on the psychological or cognitive aspects of members. Hamid et al. (2014) identified cohesion circle with limited data, such as number of mutual friends, mutual groups and common apps they are using, sometimes this mechanism works but in reality it is unable to find out psychological tie among the members.

3.2.3 Influential communities

Kafeza et al. (2013) detected influential communities based on a specific topic (i.e., #SocialNetworks) and personality traits of users. They classified personality traits into four basic categories, such as popular, energetic, conversational, and multi-systemic. Later they extended their work with Big5 model in Kafeza et al. (2014). Along with personality, they used basic Twitter metrics (e.g., tweets, retweets, hashtags, etc.) to extract influential communities. However, the results of these studies Kafeza et al. (2013, 2014) have following limitations, such as (1) Authors extracted topic-based tweets of users for a time interval. This method suffers in collecting limited features to measure personality accurately; (2) Authors extracted highly active community in a public network which possess members with strong personality profile based on Big5 (even if two members are completely disconnected from each other). However, in this paper we are interested in finding community of an ego who are also likely to be same

in the real world (e.g., Facebook). (3) It is unclear whether the members inside identified communities are similar in their behavioral aspects. Authors' in Kafeza et al. (2014) mainly focus on identifying influential community where they consider personality traits as an additional parameter. Since authors identify community in an open network, the members may be disconnected in real life. It remains unclear to what extent the members inside the identified communities are similar to the real-world situations. In our work, we investigate whether members are likely to possess similar personality traits in the real world if they are discovered in the same homophily. We introduce a novel statistical reliability analysis technique to show a high similarity among the homophily members. Note that, both Kafeza et al. (2014) and our work consider personality trait while identifying a community.

In light of the above discussion, we find that our approach differs from the existing studies in different aspects. None of these approaches discovered personality trait-based homophily with ensemble of multiple personality identification techniques in an egocentric network, which is one of the main goals of our work. Our framework is able to discover five different homophilies based on five personality traits. We devise a novel interview-based technique to evaluate the strength of our homophily structure. To the best of our knowledge, our approach is the first such innovative technique to asses homophily. In this research, we have bridged the gap between personality trait-based homophily formation and approval of homophily members in the real world.

4 Data collection

We have invited 865 users to collect Facebook statuses through posts on Facebook, relevant mailing lists and word of mouth technique. Since Facebook is a closed network, we collect *statuses* of a community that are known to each other. In our experiment, we use *judgmental sampling* technique Marshall (1996), because we first identify most productive Facebook friends who might response in our survey actively. Later, we create a Facebook application that accesses to the users' status updates. Among the 850 Facebook users, 663 members (male = 380, female = 283) agreed to share their data through the application. The rest 202 Facebook users have not shown interest to share their time-line through the application. The users are members of university student and professional community, and aged between 18 and 42 years. We build a representative dataset from different age groups and professions. We collect only 96,751 Facebook English statuses as of July 25, 2016. Maximum, minimum and average word counts of the collected statuses are 6786, 145 and 854.73.

We have conducted 44 item IPIP John (2000) test among these 663 users to collect ground truth data on Big5 personality scores. The users are asked to fill out the survey questionnaire via an experimental web page. We have also collected a new dataset of 155 (male = 97, female = 58) Facebook users who are known to each other to validate homophilies in real life. We have collected another dataset of 123 Facebook users (male = 70, female = 53) for movie preference group recommendation based on similar personality traits of an ego.

We have collected 663 users' personality score by IPIP John (2000) test. Average scores and standard deviation of these users on personality test are shown in Table 1.

5 Personality building models

We identify personality with two different strategies: (1) with closed vocabulary (i.e., LIWC)-based approach Golbeck et al. (2011) and (2) with open vocabulary (i.e., MEH)-based approach Schwartz et al. (2013). Authors in Schwartz et al. (2013) show that closed vocabulary (i.e., LIWC)-based approach applies fixed priori of words to analyze text. LIWC only captures dictionary words (4500 words, and word stems) and ignore a large amount of words those might be important signal to analyze one's personality accurately. For example, words such as selfie, Facebook, inbox, etc., are not analyzed by LIWC that are also useful cues to predict one's personality. Motivated by the work Schwartz et al. (2013), we also apply open vocabulary-based approach in our dataset of 663 Facebook users. First, we build personality identification model by using LIWC Golbeck et al. (2011). Later, we also build personality identification model using the open vocabulary approach. Then, we compare the strength between closed Golbeck et al. (2011) and open Schwartz et al. (2013) vocabulary-based approaches to show which one better predict personality scores. Finally, we select a model that predict better personality scores between two approaches.

Table 1 Average Big5 IPIP personality score on a normalized 0–1 scale

	Open.	Consc.	Extra.	Agree.	Neuro.
Average	0.6597	0.6271	0.5932	0.6515	0.4971
SD	0.1213	0.1890	0.1257	0.1137	0.1984

5.1 Closed vocabulary-based approach

In this subsection, we conduct our experiment with extensively used closed vocabulary-based approach. For building personality prediction model, we consider IPIP test result of 663 users as the ground truth data of personality traits. Motivated by the prior work on *personality* prediction from Golbeck et al. (2011), we measure word uses in users status updates with LIWC. LIWC 2007 determines 74 different types of categories, each contains hundred of words Pennebaker et al. (2007). We exclude the categories that are non semantic (e.g., proportion of long words, and filler).

We calculate Pearson correlation analysis between Big5 scores and each of the score of LIWC features. We conduct the correlation analysis among statuses of total 663 users who attended in the IPIP test. We analyze the association through linear regression to predict the score of a given personality score. We find that a number of LIWC features are correlated with a personality dimension. A potential problem arises when collinearity found between personality and LIWC features. When there is a perfect linear relationship exists among independent variables, the outcome for a regression model cannot be unique. We check variance inflation factor (VIF) among the independent variables to detect collinearity problem Chen et al. (2014). To remove collinearity among independent LIWC features, we have computed lasso penalized linear regression using *glmnet* R package Chen et al. (2014), Hastie and Qian (2014). This technique reduces the coefficients to a low value or zero, thus the model does not get overfitted. Table 2 presents that openness personality trait has the strongest (24.4 %) and neuroticism personality trait has the weakest (16.2 %) strength among all the personality traits. We find that these models moderately fitted across all the personality traits. The result has a low relative error (MAE are ranged from 0.091 to 0.127) which indicates that the model performs better than the constant mean baseline.

Motivated by the work Chen et al. (2014), we also investigate the prediction potential using a machine learning classification study. Sumner et al. (2012) suggested that computing MAE and RMSE for error measure in regression analysis is not adequate. In particular, when the majority of the individuals are around the mean of unimodal distribution, these error measures can often mask large errors.

According to the suggestion of Sumner et al. (2012), Chen et al. (2014), we apply different supervised binary machine learning algorithms on our dataset. We classify above-median level as *high* class label and below-median as *low* class label value dimension. We have experimented with few classifiers including Logistic Regression, Naive Bayesian, Adaboost, Random Forest, support vector machine and RepTree classifiers using WEKA Hall et al. (2009) machine learning toolkit. For each personality trait, we have applied these classifiers to understand the prediction performance of these personality building models.

Table 3 presents the best classifier, content type, its true positive rate (TPR), true negative rate (TNR) and Area under the ROC curve (AUC) for computing each of the personality traits Fawcett (2006). Performance of the classifiers were conducted using AUC values under the tenfold cross validation. The curve is plotted the TPR against the FPR at different threshold. The space of ROC curve is better than another if it is to the northwest (tp rate is higher, fp rate is lower, or both) of the first Fawcett (2006). We observe that our classifiers achieved moderate improvement over random chances for openness, extraversion and agreeable personality traits. For the rest of personality traits (i.e., conscientiousness and neuroticism), our models achieved lower potential than random chances.

5.2 Open vocabulary-based approach

We again analyze our dataset with open vocabulary-based approach using MEH Boyd (2014). We analyze two categories of words: (1) words ('1-g') and (2) topics.

First, we analyze '1-g' words with a big data analysis tool, MEH. Unlike content coding software (e.g., LIWC), the MEH is highly dynamic for extracting words and phrases from a dataset. Motivated by the work Schwartz et al. (2013), we extract two different types of linguistic features: (1) '1-g' words, and (2) topics. During analysis of '1-g' words with MEH, we use *MEH-Output_verbose* file which contains frequency of '1-g' words, represented as percentage of each observation. We find a total of 803 unique '1-g' words excluding stop words.

Motivated by the study Schwartz et al. (2013), we also compute another type of language feature, *topics*, consists of word using Latent Dirichlet Allocation (LDA) Blei et al. (2003). We use R *Mallet* package implementation Hornik and Grün (2011) to extract top 267 frequent topics and their percentage of frequency from our dataset of 663 Facebook users.

Table 2 Adjusted R^2 scores of the linear regression models

Big5 traits	Open. (%)	Consc. (%)	Extra. (%)	Agree. (%)	Neuro. (%)
Adjusted R^2	24.4	17.6	23.9	21.2	16.2

Table 3 Best performing classifier to predict different traits of personality using LIWC

Big5 traits	Highest AUC achieving classifier	AUC	TPR	TNR
Open.	Logistic reg.	0.617	0.669	0.472
Consc.	Logistic reg.	0.569	0.544	0.519
Extra.	Naive Bayes	0.605	0.631	0.402
Agree.	Logistic reg.	0.602	0.491	0.424
Neuro.	Logistic reg.	0.587	0.593	0.398

Then, we compute Pearson correlation between percentage of words and IPIP test result for both '1-g' words and topics, independently.

Unlike LIWC, Open vocabulary-based approach does not categorize similar words into a group. Thus, similar words contribute to build the model individually each time, which might generate collinearity among independent variables. Open vocabulary-based approach considers similar words (i.e., happy, cheer and joy) as an individual predictor. Thus, we identify a large number of collinear independent variables by observing VIF scores. Later, we compute *lasso penalized linear regression* using *glmnet* R package Chen et al. (2014), Hastie and Qian (2014) to remove collinearity among independent '1-g' words and topics. This technique reduces the coefficients to a low value or zero, thus the model does not get overfitted. We use linear regression algorithm each with a tenfold cross validation with 10 iterations. Table 4 presents the adjusted R^2 strength across all the personality traits for each type of linguistic feature.

Sumner et al. Sumner et al. (2012) suggested that computing MAE and RMSE is not sufficient to check the prediction potential of a regression model. Using the similar approach that used in Sect. 5.1, we compute prediction potential by different supervised binary machine learning algorithms on our dataset. For each personality trait and each type of linguistic features (i.e., '1-g' words and topics), we have applied previous classifiers to understand the prediction performance of these personality building models. Tables 5 and 6 present best performing classifiers to predict trait of personality using '1-g' words and topics, respectively. Authors Schwartz et al. (2013) did not demonstrate their analysis by removing collinear variables. They also did not show the prediction potential of their model using classification technique. Since classification with cross validation is reliable metric to assess how well a model work for unseen data Sumner et al. (2012), we have investigated the potential of our models with classification techniques. We observe that these classifiers achieved significant improvement than random chances.

6 Model Selection

Authors Schwartz et al. (2013) describe that closed vocabulary-based approach (i.e., LIWC) suffers less expressiveness due to a priori fixed set of words (e.g., 4500 dictionary words, word stems and 74 categories). Apart from dictionary words, SNS users generally use diverse set of words e.g., local dialects, emoticons, buzz words (i.e., selfie). LIWC-based approach does not consider these words while words being analyzed. Thus, we may ignore a large portion of personality cues while these texts being analyzed. In contrast authors Schwartz et al. (2013) proposed data driven approach, where they consider all the words (both dictionary and non-dictionary) are written by users. Open vocabulary-based approach suffers severe collinearity problem due to a large number of independent variables. We solve the collinearity problem in the Sect. 5.2 that was ignored in previous study Schwartz et al. (2013).

LIWC only analyzes word categories, and it does not capture phrase level words (i.e., '2-g' words, '3-g' words, etc). Since, we compare between open and closed vocabulary-based approaches, we ignore phrase level words in open vocabulary-based approach. Thus, we use '1-g' words and topics in our dataset to compute personality building models using open vocabulary-based approach. Tables 2 and 4 present the strength of personality building model between open and closed vocabulary-based approaches. We observe that closed vocabulary-based approach shows moderate prediction strength across all personality traits. Conscientiousness and neuroticism personality building models show low prediction strength across all the personality traits. On the other hand, open vocabulary-based approach shows better

Table 4 Adjusted R^2 scores of the linear regression models using '1-g' words and topics

Big5 traits	Open. (%)	Consc. (%)	Extra. (%)	Agree. (%)	Neuro. (%)
1-g	64.1	38.6	57.1	59	40.3
Topic	49	45	42	47.1	46.2

Table 5 Best performing classifier to predict different traits of personality using *1-g* words

Big5 traits	Highest AUC achieving classifier	AUC	TPR	TNR
Open.	RandomTree	0.669	0.671	0.461
Consc.	SVM	0.589	0.653	0.451
Extra.	RepTree	0.653	0.673	0.536
Agree.	Adaboost	0.659	0.66	0.48
Neuro.	SVM	0.597	0.602	0.493

Table 6 Best performing classifier to predict different traits of personality using *topics*

Big5 traits	Highest AUC achieving classifier	AUC	TPR	TNR
Open.	SVM	0.621	0.653	0.461
Consc.	RepTree	0.625	0.601	0.443
Extra.	RepTree	0.601	0.579	0.531
Agree.	Adaboost	0.623	0.601	0.542
Neuro.	SVM	0.638	0.669	0.391

prediction strength (see Table 4) than closed vocabulary-based approach.

Based on the personality prediction strength from Tables 2, 3, 4 and 5, we finally select open vocabulary approach as our working model. We observe in open vocabulary-based approach that one linguistic feature predicts better personality prediction score than other features for few personality traits. For example, openness, extraversion and agreeableness personality traits using '1-g'-based model shows better prediction potential than topics-based model. Again, topical modeling linguistic feature shows better prediction strength for conscientiousness and neuroticism personality traits than '1-g' linguistic feature. Since every linguistic feature contributes to compute the personality score based on their strength (weaker or stronger), to find out final personality score, we combine all the linguistic features obtained from the previous steps. It is observed in previous studies Jiménez (1998) and Polikar (2006) that combining different experts, we can build better model to predict an attribute (i.e., personality).

6.1 Ensemble of models

It is necessary to prioritize the features based on their importance, as we compute personality scores from two linguistic features (i.e., '1-g' words and topics) in Facebook. For example, some may think that '1-g' feature can reveal a personality score of a person more accurately, while other may emphasize on 'topics' to determine the personality score correctly. Ordering among linguistic features associates different weights to compute final personality score. Weight signifies the relative importance of a particular linguistic feature type. To build our ensemble/combined model, we perform following two steps: (1) computing weights from neural networks, (2) combining the personality building model with a weighted linear ensemble technique.

6.1.1 Learning weights from neural networks

In this subsection, we determine the *weight* of each linguistic feature type (e.g., '1-g' words and topics) to determine personality trait. For each type of linguistic feature and each personality traits, we model a neural network with a new dataset of 198 Facebook users (30 % of our total dataset). We model our network with two types of linguistic features and five types of personality traits; we build in a total of 10 (2×5) neural networks using R *caret* package implementation Kuhn (2008). For a single neural network, we use nine input neurons in the input layer, five neurons in the first hidden layer, three neurons in the second hidden layer and one output neuron in the output layer. For each personality trait, we take linguistic scores (i.e., percentage of *happy* '1-g' words) as input and gives a *personality* prediction score as output.

Consider a scenario, where we are interested in predicting personality score openness for the linguistic feature *topics*. We select the best subset of linguistic features using R *leaps* package implementation Lumley and Miller (2009) by forward selection approach. Then, we normalize the linguistic feature scores (i.e., topics) in the interval [0,1] with max-min normalization technique to get better precision. We keep 90 % data points of new dataset in the training set and the rest are in the test set using tenfold cross validation with 10 iterations. For each feature type and personality, we compute the strength of different models. Table 7 presents the strength (the adjusted R^2) of our neural network-based linear regression models that will be used as *weights* of our ensemble models in the next Sect. 6.1.2.

Table 7 Weights (the adjusted R^2) derived from neural networks

Big5 traits	Adjusted R^2 score of '1-g' words (%)	Adjusted R^2 score of topics (%)
Open.	25.2	21.5
Consc.	11	17
Extra.	20	18
Agree.	17.3	14.2
Neuro.	16.1	21.5

6.1.2 Weighted linear ensemble

In this subsubsection, we build a weighted linear ensemble model from different types of features of 663 Facebook users Sill et al. (2009). We have already built different models from '1-g' words and *topics* linguistic features that are described in Sect. 5.2. Since we train different neural networks that produce weights, we compute weighted linear ensemble score using the weights in Table 7.

Finally, we build our weighted linear ensemble model using the weights generated from another dataset (according to Table 7), thus our models do not get over-fitted. Table 8 presents the strength of our ensemble models and performance of the respective classifiers. We observe that our models obtain a substantial improvement with prediction potential compared with single feature-based personality identification models (according to Table 2).

Note that, we use two different datasets for our weight learning and training. Using two different datasets is somewhat similar to cross validation where we learn from one dataset and apply on another dataset. If we learn weights (i.e., contribution of different content type) from dataset and then again apply the ensemble on the same dataset, this would be like doing training/testing on the same dataset. Thus, we keep the training and testing dataset separate while building ensemble.

7 Homophily identification

In this section, first we have done MEH analysis on the users' statuses. Then, we collect Big5 scores of the 663 users from 44 item IPIP test. Later, we compute lasso penalized linear regression between two linguistic features (i.e., '1-g' words and topics) extracted from users' statuses and Big5 scores of the users using *glmnet* R package Chen et al. (2014), Hastie and Qian (2014). Then, we build ensemble-based personality prediction model that we have described in Sect. 6.1. After building the personality model, we predict Big5 personality scores over the new dataset of 155 users. Later we identify clusters of similar individuals with respect to an ego. Figure 2 shows our methodology to identify a personality trait-based homophily:

The users are the members of a student community. The members are aged between 20–35 years, and have a diverse education majors. If the personality score of a user i is similar to *ego* u, then we denote the user as *alter* i. To extract similar alters of an ego, we apply agglomerative hierarchical clustering Murtagh and Contreras (2012) technique. We extract different clusters with different stretch values. We find the best clustering result at a heuristic stretch value of $+0.02$ to -0.02. We claim that members within same cluster possess similar personality trait. For each personality trait of an ego u, we build five separate homophily clusters. Alter v_i may belong to different clusters of an ego u. This indicates that alter v_i has close match with the ego u from multiple personality traits. Our models for five personality trait-based homophilies correctly clustered users ranged from 73 to 87 % among these 155 users. We find that openness and neuroticism personality traits cluster users with an accuracy highest 87 % and lowest 73 %, respectively.

8 Homophily Validation

In this section, we validate whether homophily members of an ego are similar to each other in real life that we have extracted in Sect. 7. Every ego contains five homophilies based on five personality traits. To validate these homophilies, we ask individuals about themselves and other members who belong to same homophily about their personality using the trait-based IPIP test in real life. We divide the 44 item questionnaire into five sets which we call *trait-based IPIP questionnaire*. Finally, we check the reliability of the answers reported by homophily members using statistical techniques (e.g., ICC and Cohen's Kappa). Figure 3 shows our methodology to validate multiple personality trait-based homophilies.

Table 8 Adjusted R^2 scores of the ensemble models

Big5 traits	Open. (%)	Consc. (%)	Extra. (%)	Agree. (%)	Neuro. (%)
Adjusted R^2	75	66.9	70.3	73	65.8

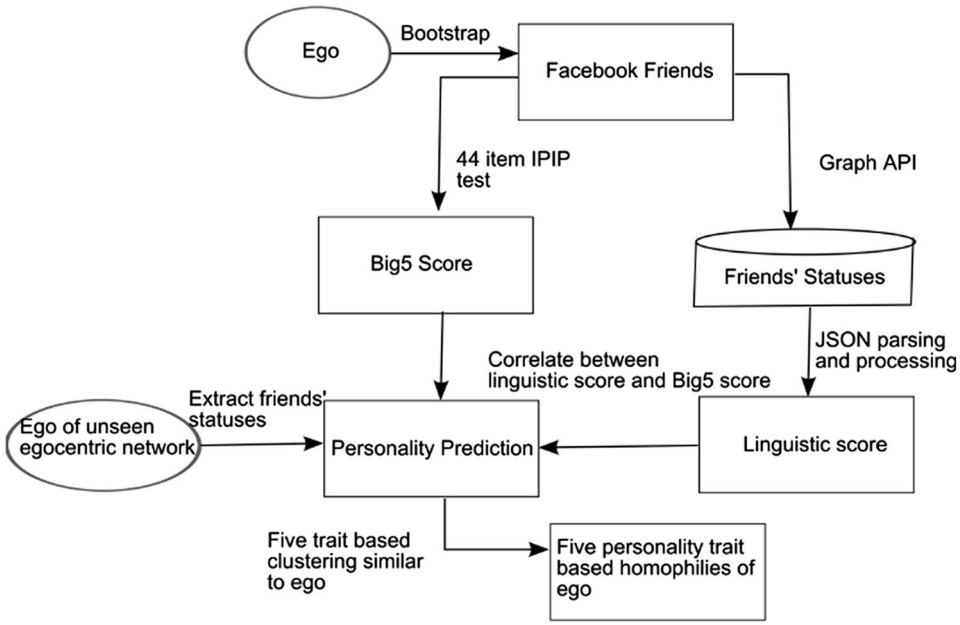

Fig. 2 Methodology of personality trait-based homophily identification

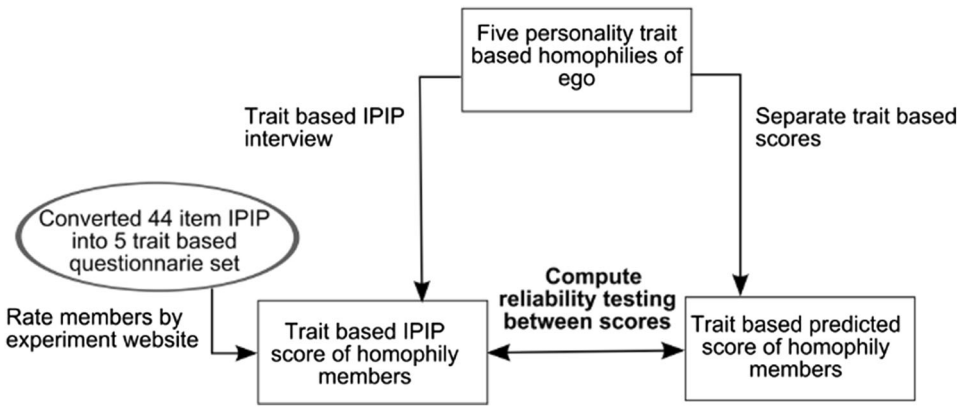

Fig. 3 Methodology of personality trait-based homophily validation

8.1 Individual trait-based IPIP scoring method

We prepare an experimental web page where we take the trait-based IPIP test. Trait-based IPIP test has five sets of questionnaires for openness, conscientiousness, extraversion, agreeableness and neuroticism, respectively. Each trait has 8–10 questions based on the trait. For example, when we test extraversion trait, we ask questions 1, 6R, 11, 16, 21R, 26, 31R, and 36 from 44 item IPIP questionnaire pool following the procedure mentioned in John et al. (2008). Among the questions, some are marked as R (reverse). For each of the question that is marked as *R* for a trait, we reverse the score in a scale of 1-5, e.g., in likert scale 2 is considered as 4 for an *R* marked question. Then for each trait, we compute max-min normalization to get the final score out of [0-1]. We test our method with the proven IPIP sitesJohn (2000) and find identical result on different personality traits.

8.2 Validation results

In this section, we present experimental results of our proposed personality trait-based homophily validation technique. Though people use IPIP test to self report themselves, but researchers also show that individual can check his similarity with his friends, co-workers, and others John (2000). Hence, we conduct experiment where individual reports answer to questions about himself as well as about his friends. In our approach, if a user is discovered in the homophily of an ego from the homophily identification predicted score, he is directed to the experimental Web site to fill up the personality trait-based IPIP questionnaire. For different users, homophily network might be different. For example, we randomly pick a member as an ego from our new dataset of 155 members. We then select 45 distinct members by hierarchical clustering technique for different personality trait-based predicted scores (Open.-22, Conc.-

13, Extra.-15, Agree.-11, Neuro.-10) who are similar to the ego. Within 45 selected members, 9 members fall into multiple trait-based homophilies of the ego. Raters and ego need a common set of known friends inside a homophily to get actual rating based on real-world relationship. Since every person rates other members in the homophily, we need a measurement of their scoring reliability.

8.3 Reliability among scores of all raters in a homophily

We compute ICC to assess the strength of consensus among the internal raters. Inter-rater reliability describes what percent of our ratings are real Koch (1983). We use two-way random model. We compute *consistency* and *absolute agreement* analysis of the ratings by the homophily members. For consistency, raters need not to agree perfectly among themselves. If their changing behavior of ratings is same, then they possess high consistency score. Later, we also compute absolute agreement. For absolute agreement, raters need to agree perfectly about their ratings. For both of these cases, we compute single and average measurement. Single measurement determines to what extent rating of a single person is reliable, if he rates himself. Average measurement determines reliability of raters on average. Table 10 compares between *ICC* scores of personality trait based and two baseline homophily identification techniques.

We also compute *Kappa* Viera et al. (2005) by the observation of two external raters among the members of the homophily. Though these two raters are not a homophily member, they need to know all of the members of a homophily to rate them accurately. Raters judge these homophily members with particularly low (0–0.5) or high (0.51–1.0) values of a trait. Table 11 compares *kappa* values between personality trait based and two other baseline homophily identification techniques.

9 Comparison with other baseline techniques

In this section, we propose and justify two comparable baseline techniques with our proposed technique to identify personality trait-based homophilies. A number of user activities such as tagging items and listening music have been identified to extract homophilies Aiello et al. (2012), Bisgin et al. (2010). In this paper, we propose two baseline techniques to identify homophilies are: (1) users who like similar Facebook fan pages, and (2) users who interact frequently in Facebook. We first find out members of homophilies who are similar to an ego in terms of two different criteria (e.g., *page-likes* and *degree of interaction*). Then, we assume that these homophily members are likely to possess similar personality trait of an ego as they behave similarly such as *liking* similar pages and interacting frequently with similar members in Facebook. To validate these homophilies, we apply two baseline homophily identification techniques with our evaluation dataset of 155 Facebook users. Later, we ask these three different homophily members (personality trait-based homophily with two baseline techniques) about the similarity of their personality by trait-based *IPIP* questions (according to Sect. 8.1). Finally, we show that our technique more accurately distinguish personality trait-based homophilies than that of two baseline techniques.

9.1 Page-likes similarity

In this baseline technique, we identify homophilies of Facebook users who *like* similar Facebook fan pages of an ego. We consider that users who *like* similar Facebook Fan pages are likely to possess similar personality traits. For example, a person with high score in openness is likely to *like* pages that present content with excitement, new experiences and strong stimulus. On the other hand, a person with strong conscientiousness personality score usually like pages with career building, health awareness, etc. Thus, we identify homophily (similar) members among the 155 users of evaluation dataset for Facebook page-like similarity. We collect *page-likes* (tag: 'about') of these users using our Facebook application. We collect a total of 19,766 page-likes (tag: "about") as of July 25, 2016 through our Facebook application. The *page-likes* dataset has a maximum, minimum and average word counts 1887.3 and 1234.12, respectively.

Authors in Liben-Nowell and Kleinberg (2007) predicted links based on similarity of co-authorship network. Following their technique, we also identify homophily of an ego based on their patterns of *liking* Facebook pages. We consider all of the *page-likes* of a single user as a single document. We compute *bag-of-words* among the documents of all users by removing stop words, and using lemmatization technique. Given an ego u, we compute similar alters V' based on similarity of *terms* found in user *page-likes* document using Jaccard's coefficient. Let us consider that $X = (x_1, x_2, \ldots, x_n)$ and $Y = (y_1, y_2, \ldots, y_n)$ are two vectors of term frequency of page-likes of an ego u and an alter $v_j \in V'$, respectively. Then, we can compute the Jaccard's coefficient by the following equation:

$$J(x_i, y_i) = \frac{\sum_i |\Gamma(x) \cap \Gamma(y)|}{\sum_i |\Gamma(x) \cup \Gamma(y)|} \quad (1)$$

In this baseline, we compute Jaccard's coefficient between *page-likes* documents of an ego u and all the alters $u' \in V'$. After computing the Jaccard's coefficients, we apply agglomerative hierarchical clustering Murtagh and

Contreras (2012) that we used during homophily identification in Sect. 7. To make an unbiased experiment, we keep same stretch value of +0.02 to D. In our clustering result, we find 45 alters belong to the same cluster (homophily) of the ego u.

We assume that these 45 *alters* possess similar personality traits of ego u, since they like similar Fan pages to ego in Facebook. To prove our hypotheses, we conduct experiment where these alters report answer to IPIP test about himself as well as about his members in the same homophily. Tables 10 and 11 shows that *ICC* agreement and *Kappa* coefficient in different personality traits are *poor* based on *page-likes* homophily result. Thus, we conclude that Facebook users who *like* similar pages do not possess similar personality traits in real world.

9.2 Maximum degree of interaction

In this baseline technique, we identify homophilies among users who frequently interact among themselves in Facebook. Authors in a study Dev et al. (2014) extract *active communities* by computing the *degree of interactions* among users. Higher degree of interaction signifies that strong tie strength exists among the users and they are similar with each other than the rest of members in the network. Based on the interactions, it is possible to find out a group of users (homophily). We call this homophily as *interaction-based homophily*. We assume that these users have similarity among themselves, since they have strong tie strength and they interact frequently in Facebook. Motivated by the approach Dev et al. (2014), we identify $u' \in V'$ users (alters) who have highest degree of interactions with ego u. We call that these users (alters) belong to the same homophily of ego u.

To extract interaction graph of Facebook ego network, we use NodeXL Hansen et al. (2010), an open-source network analysis and visualization tool. We collect interaction data such as *likes, comments*, and *sharing* of object between an ego u, and each alter u'. We consider the same 155 alters those we used in previous experiments. We count total degree (sum of in-degree and out-degree) for each pair of ego u and alter connection. For each interaction (e.g., commenting or liking a post) between them, we increase the total count by one. We collect all the interaction data between ego u and each alter u' as of July 25, 2016 through NodeXL tool. We find maximum, minimum and average interaction degree counts are 183.7 and 25.31, respectively.

Then, we build homophily (cluster) of ego u based on higher degree of interactions. We select the cluster that contains 45 different alters, since the size of our *page-like* homophily is also 45. For the sake of uniform clustering parameters in all experiments, we keep the size of cluster same. We assume that these 45 users possess similar personality traits as they have higher degree of interactions with the ego u. To prove our hypotheses, we conduct experiment (the same experiment that we conduced in Sects. 7 and 9.1) where these alters report answer to *IPIP* test about himself as well as about his members in the same cluster. Tables 10 and 11 show that *ICC* agreement and *Kappa* coefficient in different personality traits are *poor* based on *degree of interaction* result. Thus, we infer that users who interact frequently through the Facebook do not possess similar personality traits in real world.

10 Personality trait-based group recommendation

Group recommendation is a well-studied topic in recommender systems Amer-Yahia et al. (2009), Gorla et al. (2013). Recommending a group has a number of real scenarios that include selecting restaurants for taking lunch with friends, finding travel destination with family members, etc. In this paper, we show that finding personality trait-based homophilies can facilitate to recommend members of a group to perform certain task by an ego. In particular, we demonstrate how to find out the preferences of movies for a group of Facebook users who are similar to ego u based on their personality traits.

It is evident from previous studies Feng and Qian (2013), Chen et al. (2013) that personality traits influence user needs and preferences. Motivated by the prior studies Adamopoulos and Todri (2015), Hsieh et al. (2014), we investigate whether there is any correlation between user personality traits and group movie preference of users in Facebook. The intuition behind our movie preference scenario is as follows. Since individuals possess different personalities, it is highly likely that the preference of movie watching may differ based on the type of movie contents. In this study, we investigate whether we can accurately recommend a group of users to recommend movies who are likely to possess similar personality trait of ego u. We find strong correlation between personality score of ego and preference of movies of her homophily members.

10.1 Movie preference

To evaluate our models for movie preference application, we first collect a new dataset of 123 (male=70, female=53) active Facebook users through our application. Users are from the same ethnic group, but they are from different educational and professional background, aged between 20 and 28 years. Later, we predict homophily members by our models and prepare a questionnaire to find out movie preference of these users through interviews. All the recruited participants must have watched some selected

movies to attend in the interview. We recruit these participants by observing their Facebook *movie watched list*. Then, we compute correlation coefficient between the predicted personality group scores of homophily members and results of the questionnaire in real life. In particular, we will investigate the following three hypotheses that correlate personality traits with movie preferences of users.

H1 High *openness* personality homophily members are strongly associated with sci-fi/adventurous movies. An individual who possesses high score in openness personality is likely to give high importance on futuristic view and active imagination. He likes to experience challenges and excitements in his life. Since sci-fi/adventurous movies contain new ideas and stimuli in the movie contents, it is likely that an individual with high score in *openness* personality prefers to watch those movies.

H2 High *extraversion* personality is strongly associated with comedy movies. Extraversion is about enjoying life, seeking happiness, and sensuous gratification for oneself. Since comedy movies contain these attributes in their movie content, we hypothesize a positive link between extraversion personality and content of the comedy movies.

H3 High *agreeableness* personality is strongly associated with romance or drama movies. A person with high agreeable score generally trusting, generous and helpful. Since drama and romance genre of movies contain sentimental, emotional, sacrificing and compassionating content, we hypothesize a positive link between agreeableness personality score and content of romance and drama genre of movies.

10.2 Movie preference experiment

First of all, we predict homophily of ego u based on Big5 personality traits in our dataset of 123 Facebook users. In our predicted homophilies of ego u, we find 37, 22, 41, 32, and 12 homophily members in openness, conscientiousness, extraversion, agreeableness, and neurotic personality traits, respectively. According to our hypotheses, we assume that homophily members of openness, extraversion and agreeable personality traits have association with sci-fi/adventurous, comedy, and romantic/dramatic movies, respectively. To prove our hypotheses with group recommendation, we conduct the following experiments.

We conduct a semi-structured interview during July 2016 in different locations (e.g., restaurants, university library, etc.). Most of the interviews are taken in the face-to-face settings. At the end of the interview, the homophily members are compensated by a small gift. We also take few interviews through Skype for homophily members who stay in distant locations.

We initially hypothesized that the homophily members who like to watch *sci-fi/adventurous*, *comedy*, and *romance/drama* genre of movies, they tend to possess the high score in *openness*, *extraversion*, and *agreeableness* personality traits, respectively. First, we select a list of total six movies from three different genres: sci-fi/adventurous, comedy, and romance/drama. All of the movies are rated between 8.2 to 8.8 according to imdb (Internet movie database) rating. We confirm that all the homophily members have watched all of the selected movies previously. They are asked to rate all the six movies based on their preferences in a likert scale of 1–5. In likert scale, 1 is *strongly disinterested*, 2 is *disinterested*, 3 is *neither disinterested nor interested*, 4 is *interested* and 5 is *strongly interested* to recommend as worth watching a particular movie. Based on the answer of the likert scale, we normalize the score between 0 and 1. After rating these genre-based movies by the homophily members, we compute mean of normalized (0-1) score. Then, we compute correlation coefficients between mean normalized ratings of the three genre-based movies and predicted personality scores from our model for each homophily member. Table 12 presents the correlation coefficients between movie genre preferences and personality scores of Facebook users in different homophilies.

10.3 Experimental result

To evaluate our hypotheses, we observe the correlations between individuals' personality ratings with their self-reported movie preferences. Table 12 shows that **H1**, **H2**, and **H3** conform with the preference of movie genre selection in real world. We find strong correlation (0.31**) between openness personality and preference of sci-fi/adventurous movie genres. We also find strong correlation (0.281**) between extraversion personality score and preference of comedy movie genre. Similarly, we find strong correlation (0.27**) between agreeableness personality score and preference of romance/drama genre of movie.

Thus, our personality trait-based homophily identification technique is able to recommend a group of Facebook users who possess similar personality score to ego u. This technique can also recommend a group of similar users to ego u more accurately about their product, preference of gadget selection, and other similar types of applications in real life.

11 Discussion

Our work is the first study (1) to identify the correlation of word usage with personality trait-based homophily in an egocentric social network using two different personality

Table 9 Best performing classifier to predict different traits of personality using ensemble of models

Big5 traits	Highest AUC achieving classifier	AUC	TPR	TNR
Open.	RandomTree	0.743	0.772	0.335
Consc.	Logistic Reg.	0.695	0.722	0.371
Extra.	Random Forest	0.683	0.681	0.324
Agree.	Adaboost	0.713	0.704	0.327
Neuro.	Adaboost	0.679	0.67	0.341

Table 10 ICC score for different techniques of homophily computation

Homophily technique	Type	Measure	Open.	Consc.	Extra.	Agree.	Neuro.
Big5	Consistency	Single	0.619	0.253	0.589	0.539	0.231
	Absolute agreement	Average	0.872	0.501	0.807	0.788	0.457
Page-likes similarity	Consistency	Single	−0.11	0.15	0.07	0.18	0.08
	Absolute agreement	Average	0.05	0.17	0.09	0.14	−0.03
Similar k-users of interaction	Consistency	Single	0.141	0.09	0.03	0.11	−0.01
	Absolute agreement	Average	0.173	0.11	0.013	0.16	0.01

Table 11 Kappa score for different techniques of homophily computation

Homophily technique	Kappa	Open.	Consc.	Extra.	Agree.	Neuro.
Big5	Value	**0.709**	0.231	**0.632**	**0.484**	0.207
	Approx sig.	**0.001**	0.117	**0.001**	**0.002**	0.171
Page-likes similarity	Value	0.087	0.043	0.204	0.10	0.15
	Approx sig.	0.731	0.764	0.125	0.317	0.231
Similar k-users of interaction	Value	0.08	0.04	0.10	−0.01	0.10
	Approx sig.	0.734	0.79	0.343	0.831	0.31

Bold indicates openness, extraversion and agreeableness personality traits which are statistically significant while computing personality trait based homophilies

identification techniques, and (2) to validate the identified homophilies in real world.

From Tables 2 and 4, we observe that both of the techniques of open vocabulary-based approach (i.e., '1-g' words and topics) outperform closed vocabulary-based approach. In open vocabulary-based approach, openness and conscientiousness personality traits show the strongest and weakest strength, respectively, among all the personality traits using '1-g' technique. From Table 4, we also notice that conscientiousness and neuroticism personality traits show improvement using topical modeling than '1-g' technique. Again, we observe from Tables 3, 5 and 6 that open vocabulary-based approach shows moderate improvement over random chances for all personality traits than closed vocabulary-based approach. From Tables 8 and 9, we observe that our ensemble-based method shows substantial improvement over individual closed and open vocabulary-based approach.

While validating personality trait-based homophilies in real world, we find in Tables 10 and 11 that our ensemble-based technique discovered homophilies using Big5 technique show better agreement scores than other two baseline homophily identification techniques. For example, ICC and kappa values of openness homophily (Big5) show substantial (0.61–0.80) agreement among the raters Viera et al. (2005). Again, we observe that ICC and kappa values of agreeable homophily (Big5) show moderate (0.41–0.60) agreements. We also notice that extraversion homophily (Big5) shows moderate (0.41–0.60) and substantial (0.61–0.80) ICC and kappa values, respectively. In contrast, both conscientiousness and neurotic homophilies show fair (0.21–0.40) ICC and kappa agreements among the raters. It has also been proved in previous studies that personality prediction for neuroticism is difficult Back et al. (2010). Again, conscientious people have less propensity to share information in the social media frequently Hughes et al. (2012). We notice that ICC and kappa strength of different personality trait-based homophilies are correlated with the strength of personality (according to Table 8) building model. Thus, we find fair agreement among the raters while discovering homophilies for conscientiousness and neuroticism traits.

In contrast, we find slight or *less than chance* (<0.0) agreement (according to Tables 10 and 11) among the

Table 12 Correlation coefficients (CC) between group personality score and results of movie preference interviews

Movie genre	Openness	Extraversion	Agreeableness
Sci-fi/Adventurous	0.31**	0.091	−0.03
Comedy	0.07	0.281**	0.13
Romance/Drama	0.05	0.11	0.27**

* $p < 0.05$; ** $p < 0.01$

raters using two baseline (e.g., page-like similarity and similar *k*-users of interaction) techniques. For example, we find *less than chance* ICC agreement of *page-like similarity* technique for openness and neuroticism homophilies while other three trait-based homophilies show slight agreement. For kappa values, the agreements also show random scores. We also observe *less than chance* or *slight* agreement while discovering homophilies with similar *k-users of interaction* technique. We find no correlation with personality model and its derived homophilies using these two baseline techniques. Therefore, to identify personality trait-based homophilies using ensemble technique, Big5 model outperforms other two baseline techniques.

In Sect. 10, we find strong statistical correlation between personality trait-based homophily with respect to personality of an ego and movie watching preference. Table 12 presents that sci-fi/adventurous, comedy, and romance/drama movie genres are strongly correlated with openness, extraversion, and agreeableness personality trait-based homophilies, respectively. Since we get less ICC and Kappa values for conscientiousness and neurotic personality trait-based homophilies, we do not find such association to recommend movie preference.

In a previous study, it is shown that a moderate personality prediction strength can be achieved from my Personality Celli et al. (2013) dataset ($N = 250$) with minimum and average word count of 1 and 585.004, respectively. In another well cited study Golbeck et al. (2011), authors successfully predict personality from Facebook with a sample size of 279 Facebook users. Therefore, the size ($N = 663$) of our dataset is sufficient to predict personality from social media usage. There are few datasets available for modeling personality from social media Celli et al. (2013), Schwartz et al. (2013). Since we need to validate our homophilies in real world, we build and evaluate models with our own datasets that are collected from the same demographics.

12 Conclusion

In this paper, we have presented homophily identification and validation techniques for users in social media (i.e., Facebook). To identify homophily, we have first built a prediction model that takes a user's Facebook status as input and gives the personality scores of the user as output. We have identified personality prediction models using both closed and open vocabulary-based approaches. We have also compared the strength of those models using classification and regression prediction potential techniques. We have first computed personality scores by combining two different linguistic features (i.e., '1-g' words and topics). Then by using hierarchical clustering technique, we have identified a community for an ego, where five different personality traits-based similarities are considered for the identification of homophilies. We have also presented a novel validation technique that enables us to verify and compare our identified homophilies with other techniques in real world. Our work is the first one that identifies personality trait-based homophilies from Facebook and validates those homophilies in real world. We have also demonstrated group movie recommendation, an application of personality trait-based homophily identification. In future, we plan to use *Empath* for deriving personality trait-based homophilies from multiple interaction features (i.e., page-likes, shared-links, etc) that generates on demand new lexical categories Fast et al. (2016) and compare with other approaches. We are also interested in combining our method with an existing recommendation system and investigate the overall performance.

Acknowledgments This research is funded by ICT Division, Ministry of Posts, Telecommunications and Information Technology, Government of the People's Republic of Bangladesh.

References

Adali S, Golbeck J (2012) Predicting personality with social behavior. In: ASONAM. IEEE

Adamopoulos P, Todri V (2015) Personality-based recommendations: evidence from amazon.com. In: Proceedings of the 9th ACM international conference on recommender systems

Aiello LM, Barrat A, Schifanella R, Cattuto C, Markines B, Menczer F (2012) Friendship prediction and homophily in social media. TWEB 6(2):9

Amer-Yahia S, Roy SB, Chawlat A, Das G, Yu C (2009) Group recommendation: semantics and efficiency. Proc VLDB Endow 2(1):754–765

Arnaboldi V, Conti M, Passarella A, Dunbar R (2013) Dynamics of personal social relationships in online social networks: a study on twitter. In: Proceedings of the first ACM conference on online social networks. ACM, pp 15–26

Back MD, Stopfer JM, Vazire S, Gaddis S, Schmukle SC, Egloff B, Gosling SD (2010) Facebook profiles reflect actual personality, not self-idealization. Psychol Sci 21:372

Bisgin H, Agarwal N, Xu X (2010) Investigating homophily in online social networks. In: WI-IAT. IEEE, vol 1, pp 533–536

Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. J Mach Learn Res 3:993–1022

Boyd R (2014) Meh: meaning extraction helper (version 1.0.6)

Celli F, Pianesi F, Stillwell D, Kosinski M (2013) Workshop on computational personality recognition (shared task). In: Proceedings of the workshop on computational personality recognition

Chen L, Wu W, He L (2013) How personality influences users' needs for recommendation diversity? In: CHI'13 extended abstracts on human factors in computing systems. ACM, pp 829–834

Chen J, Hsieh G, Mahmud JU, Nichols J (2014) Understanding individuals' personal values from social media word use. In: Proceedings of the 17th ACM conference on computer supported cooperative work and social computing. ACM, pp 405–414

Crandall D, Cosley D, Huttenlocher D, Kleinberg J, Suri S (2008) Feedback effects between similarity and social influence in online communities. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, pp 160–168

Dev H, Ali ME, Hashem T (2014) User interaction-based community detection in online social networks. In: DASFAA. Springer, pp 296–310

Fast E, Chen B, Bernstein M (2016) Empath: understanding topic signals in large-scale text. arXiv preprint arXiv:1602.06979

Fawcett T (2006) An introduction to roc analysis. Pattern Recognit Lett 27(8):861–874

Feng H, Qian X (2013) Recommendation via user's personality and social contextual. In: Proceedings of the 22nd ACM international conference on information and knowledge management. ACM, pp 1521–1524

Fisher D (2005) Using egocentric networks to understand communication. IEEE Internet Comput 9(5):20–28

Gilbert E, Karahalios K (2009) Predicting tie strength with social media. In: Proceedings of the SIGCHI conference on human factors in computing systems. ACM, pp 211–220

Golbeck J, Robles C, Turner K (2011) Predicting personality with social media. In: CHI'11. ACM, pp 253–262

Gorla J, Lathia N, Robertson S, Wang J (2013) Probabilistic group recommendation via information matching. In: Proceedings of the 22nd international conference on World Wide Web. ACM, pp 495–504

Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH (2009) The weka data mining software: an update. ACM SIGKDD Explor Newslett 11(1):10–18

Hamid MN, Naser MA, Hasan MK, Mahmud H (2014) A cohesion-based friend-recommendation system. Soc Netw Anal Min 4(1):1–11

Hansen D, Shneiderman B, Smith MA (2010) Analyzing social media networks with NodeXL: insights from a connected world. Morgan Kaufmann, Los Altos

Hastie T, Qian J (2014) Glmnet vignette. Technical report, Stanford

Hornik K, Grün B (2011) Topicmodels: an r package for fitting topic models. J Stat Softw 40(13):1–30

Hsieh G, Chen J, Mahmud JU, Nichols J (2014) You read what you value: understanding personal values and reading interests. In: Proceedings of the 32nd annual ACM conference on Human factors in computing systems. ACM, pp 983–986

Hughes DJ, Rowe M, Batey M, Lee A (2012) A tale of two sites: twitter vs. facebook and the personality predictors of social media usage. Comput Hum Behav 28(2):561–569

Jiménez D (1998) Dynamically weighted ensemble neural networks for classification. In: The 1998 IEEE international joint conference on neural networks proceedings, 1998. IEEE world congress on computational intelligence. IEEE, vol 1, pp 753–756

John OP (2000) The big five personality test. http://www.outofservice.com/bigfive/. Accessed 25 July 2016

John OP, Srivastava S (1999) The big five trait taxonomy: history, measurement, and theoretical perspectives. Handb Pers: Theory Res 2(1999):102–138

John OP, Naumann LP, Soto CJ (2008) Paradigm shift to the integrative big five trait taxonomy. Handb Pers: Theory Res 3:114–158

Kafeza E, Kanavos A, Makris C, Chiu D (2013) Identifying personality-based communities in social networks. In: Parsons J, Chiu D (eds) Advances in conceptual modeling. Springer, Hong Kong, pp 7–13

Kafeza E, Kanavos A, Makris C, Vikatos P (2014) T-pice: twitter personality-based influential communities extraction system. In: Parsons J, Chiu D (eds) BigData congress. IEEE, PhD Symposium, Hong Kong, pp 212–219

Koch GG (1983) Intraclass correlation coefficient. In: Encyclopedia of statistical sciences, vol 4. Wiley, pp 212–217

Kosinski M, Stillwell D, Graepel T (2013) Private traits and attributes are predictable from digital records of human behavior. Proc Natl Acad Sci 110(15):5802–5805

Kuhn M (2008) Caret package. J Stat Softw 28(5):1–26

Liben-Nowell D, Kleinberg J (2007) The link-prediction problem for social networks. J Am Soc Inf Sci Technol 58(7):1019–1031

Lumley T, Miller A (2009) Leaps: regression subset selection. R package version 2.9. See http://CRAN.R-project.org/package=leaps

Marshall MN (1996) Sampling for qualitative research. Fam Pract 13(6):522–526

McAuley JJ, Leskovec J (2012) Learning to discover social circles in ego networks. NIPS 2012:548–56

McPherson M, Smith-Lovin L, Cook JM (2001) Birds of a feather: homophily in social networks. Ann Rev Sociol 27(1):415–444

Murtagh F, Contreras P (2012) Algorithms for hierarchical clustering: an overview. Wiley Interdiscip Rev: Data Min Knowl Discov 2(1):86–97

Norman WT (1963) Toward an adequate taxonomy of personality attributes: replicated factor structure in peer nomination personality ratings. J Abnorm Soc Psychol 66(6):574

Pennebaker JW, Booth RJ, Francis ME (2007) Linguistic inquiry and word count: Liwc. Austin: liwc. net. http://www.liwc.net/LIWC2007LanguageManual.pdf. Accessed 29 July 2016

Petrocelli T (2014) Closed vs open social networks

Polikar R (2006) Ensemble-based systems in decision making. IEEE Circuits Syst Mag 6(3):21–45

Schwartz HA, Eichstaedt JC et al (2013) Personality, gender, and age in the language of social media: the open-vocabulary approach. PLoS One 8(9):e73791

Sill J, Takács G, Mackey L, Lin D (2009) Feature-weighted linear stacking. arXiv preprint arXiv:0911.0460

Sumner C, Byers A, Boochever R, Park GJ (2012) Predicting dark triad personality traits from twitter usage and a linguistic analysis of tweets. In: 2012 11th international conference on machine learning and applications (ICMLA). IEEE, vol 2, pp 386–393

Viera AJ, Garrett JM et al (2005) Understanding interobserver agreement: the kappa statistic. Fam Med 37(5):360–363

Yarkoni T (2010) Personality in 100,000 words: a large-scale analysis of personality and word use among bloggers. J Res Pers 44(3):363–373