Revealing Psycholinguistic Dimensions of Communities in Social Networks

Tushar Maheshwari, Aishwarya N. Reganti, Upendra Kumar, Tanmoy Chakraborty, Amitava Das

Indian Institute Of Information Technology Chittoor, Sri City, A.P., India

Indraprastha Institute of Information Technology Delhi (IIIT-D), India

Abstract: A community in social network is composed of individuals with similar behavior. Although there has been a plethora of work on understanding network topologies (edge density, clustering coefficient, etc.) within a community, the semantic interpretation of a community has hardly been studied. The present paper aims at understanding Personalities and Values of individuals in social communities. To this end, we collect datasets from various social media platforms (including Facebook, Twitter), which contain Values and Ethics of users. Then we design a three-fold experimental setup. First, we propose automatic models to determine Personality and Values (Values and Ethics) of individuals by analyzing the kind of language used and their conduct in social media. Secondly, various experiments are performed to understand the blend of characteristics of individuals within a social network community. Finally, we claim that the detected Personality and Values of individuals can be used further as additional node attributes to detect better community structure. To our knowledge, this is the first computational analysis to understand and predict psycholinguistic dimensions of individuals in social networks.

Index Terms—Personality, Values, social media, societal sentiment, community

1 INTRODUCTION

Detecting and analyzing dense groups or communities from social networks has attracted immense attention over the last decade. Several heuristics and algorithms were proposed to detect communities based on the topological structure of the network [1]. However, the semantic interpretation of a community, i.e., the behavior of individuals within a community has hardly been studied. This paper presents the first computational psycholinguistic study to understand the behavior of individuals forming communities in social networks. We use two psycholinguistic models – Personality and Values models, to identify the behavior of individuals. The Personality model is used to understand the characteristics or blend of characteristics at individual level, whereas the Values model is used to analyze inter-personal dynamics of societal sentiment.

The Big 5 Personality traits [2], aka the five factor model (FFM), is a widely used Personality model. The five factors are: Openness (O): A personality trait possessed by individuals who are imaginative, insightful and have wide interests; Conscientiousness (C): Refers to those who are organized, thorough, and planned; Extroversion (E): Refers to Personality of those who are talkative, energetic, and assertive; Agreeableness (A): Individuals with this Personality trait are sympathetic, kind, and affectionate; and Neuroticism (N): Individuals who are mostly tense, moody, and anxious. Big 5 model is also represented using the acronym OCEAN.

To define societal sentiment, we use the well-established “Schwartz Theory of Basic Human Values” [3], which defines ten basic and distinct personal Values – Achievement (AC): Achievers set goals and then make all possible endeavors to achieve them; Benevolence (BE): Those who tend towards being benevolent are very philanthropic, they seek to help others and provide general welfare; Conformity (CO): This category of people obey clear rules and structures; Hedonism (HE): Hedonists are those who simply enjoy themselves; Power (PO): The ability to control others is important to people who possess this value and power will be actively sought by dominating others and control over resources; Security (SE): Those who seek security value, health and safety to a greater extent than other people (perhaps because of childhood woes); Self-direction (SD): Individuals who are self-directed, enjoy being independent and are outside the control of others; Stimulation (ST): It is closely related to hedonism, nevertheless the goals are slightly different. In this case, pleasure is acquired specifically from excitement and thrill; Tradition (TR): A traditionalist respects practices of the past, doing things blindly because they are customary; Universalism (UN): Individuals who are universal, seek social justice and tolerance for all.

Schwartz, along with the identification ten basic Values, also explains how these Values are related to each other and influence one another, since individuals possessing any of the Values may also possess values in accord with another (e.g., Conformity and Tradition) or contrary with at least one other Value (e.g., Benevolence and Achievement). Such coinciding nature psychological classes makes the computational classification problem much more challenging than the typical sentiment analysis problem.

In this paper, we raise a fundamental question in order to understand the behavioural characteristics of individuals
in social communities – what are the Personality and Values classes of individuals forming communities in social networks, and how do these differ across multiple communities? To this end, three corpora have been utilised (one collected by us and two publicly available): (i) a Twitter corpus, where users are marked with their corresponding Values, (ii) a Facebook corpus containing user Personality markings [4], and (iii) another large-scale Twitter network with ground-truth communities. Three kinds of experiments are designed over these corpora. First, systems are developed to automatically determine Personality and Values of individuals by analysing the kind of language they use and behaviour in social media (Section 4). Second, a detailed analysis is conducted to understand the dynamics of societal sentiment within a community (Section 5). Finally, we claim that the detected Personality and Values of individuals can be used further as additional node attributes to detect better community structure (Section 6).

2 Related Work

The Big 5 Personality model [2] and Schwartz’ Values model [3] can be considered as a person level sentiment model and a societal sentiment model, respectively. Going one step further, we here attempt to understand the societal sentiment of groups of individuals forming communities in social networks [5].

Intensive analysis of human beliefs which also includes their values, ethics, the kind of attitudes they possess has been a cardinal research path in both Psychology and Sociology for many years now.

According to the Schwartz 10-Values model, ten basic Values denote numerous consequences and results of an individual’s role in a society [6]. The Values have also been able to successfully provide a strong and valid explanation of consumer conduct/behaviour and how these values affect it [7].

Over the recent years, there have been few initiatives to automatically identify various Personality traits of individuals from the use of language and their behaviour in social media; one such initiative was the Workshop and Shared Task on Computational Personality Recognition held in 2013, and then again in 2014 (see Section 3.2 for more details). However, to the best of our knowledge, no automatic predictive model for Schwartz’ Values has been built or tested before. In recent times, proliferation of social media and the ever-growing collection of publicly accessible web data continually provides new and exciting opportunities to study how people are thinking, behaving, and feeling, and is the main motivation of our present research.

Community boundary detection is a fundamental problem in network analysis, which has been extensively studied during last one decade (see the detailed reviews in [8]). Most popular algorithms consider only network information for community detection. However, recently there has been a claim that augmenting node and edge attributes along with the network topology increases the performance of the algorithms [9], [10]. However, all these approaches mostly consider social dimensions of individuals such as occupation, gender, age, locality etc. as node attributes for the algorithms. We are the first to explore the psycholinguistic behavior of individuals (For further details refer to our paper [11]) and augment them as additional features into the community detection algorithm to show how it helps in detecting more accurate communities from the networks (see Section 6).

3 Corpus Acquisition

At the very beginning of our research work, we ask ourselves an elementary question- a self-interrogation - If social media is a good representative of the original society that exists?, in order to understand whether Personality and Values of human beings get reflected in their social media behaviour. To the best of our knowledge, there is no prior work on understanding Values from social media behaviors. We were motivated by a similar work [12]. In this paper, our endeavour is to understand the behavioural characteristics of individuals in social communities: What are the Personality and Values classes of individuals forming communities in social networks, and how do these differ across multiple communities? To this end, three corpora have been utilised – one collected by us and two publicly available: (i) A Twitter corpus, where users are marked with their corresponding Values, (ii) A Facebook corpus containing user Personality markings [4], and (iii) Another large-scale Twitter network with ground-truth communities.

Two kinds of experiments were designed using these corpora. First, systems were developed to automatically determine Personality and Values of individuals by analysing the kind of language they use and conduct in social media. Second, a detailed analysis was conducted to understand the dynamics of societal sentiment - involving Personality and Values of individuals within a community.

3.1 Twitter Values Corpus

The standard method for any psychological data collection is through self-evaluative psychometric tests. Self-evaluations were obtained using a fifty-item long male/female version of the Portrait Values Questionnaire (PVQ). We crowd-sourced the data using the Amazon Mechanical Turk (AMT) service. A 50 item PVQ questionnaire was given to people and we requested them to: (i) attempt the PVQ questions sincerely, and (ii) provide us with their Twitter ids so that their tweets could be scraped with due permission. However, we faced a lot a challenges while collecting the data. For example, many users had set their accounts to protected, which hindered our data collection through the Twitter API. Also, we discarded quite some users as they had posted fewer than 100 tweets (according to the statistics, around 44% of the Twitter users created an account, but never sent a tweet), making them insignificant for statistical analysis. Finally, we manage to collect information of 368 unique users. The highest, lowest and average number of tweets per user was 15K, 100 and 1,608 respectively.

In addition to identifying the basic Personality traits and Schwartz Values types, both the models also explain how
beats their system with an average 0.0 of achieved F-Score (an average over all the personality traits) anticipated in this shared task, the best performing team [4]. We further enriched the Twitter id’s. This dataset has been widely used in community detection studies [10]. We used the Twitter network, released by SNAP [3].

3.3 Twitter Community Dataset

We use the Twitter network, released by SNAP [3] (nodes: 81,306, edges: 1,768,149). Users are distinguished by their Twitter id’s. This dataset has been widely used in community detection studies [10]. We further enriched the dataset by crawling the tweets of each user, required for our Personality and Values models. The original dataset had 18,021 users and 5,038 communities. We discarded all the communities with size less than five and considered 1,562 remaining ground-truth communities. We further discarded all the non-existent accounts and those users who had posted fewer than 100 tweets. We downloaded total 6,768 users’ tweet data. The highest (resp. lowest) number of tweets for a user was 3,641 (resp. 100) with an average number tweets per user being 2,406.

4 Computational Personality and Values Models

Here, we discuss various features and methods used for the automatic Personality and Values identification. We experimented with several machine Learning algorithms – Support Vector Machine (SVM), Multinomial Naïve Bayes (mNM), Simple Logistic Regression (LR), and Random Forest (RF). Among them, SVM (with linear kernel) turned out to be the best performing method. A binary SVM classifier is trained, each for a particular Personality and Values type separately.

Psycholinguistic Lexica: We use four different psycholinguistic lexicon based features: Linguistic Inquiry Word Count (LIWC) [4], Harvard General Inquirer [5], MRC psycholinguistic database [6], and Sensicon [7]. LIWC is a diligently developed and popular hand-crafted lexicon. It contains 69 different categorical words, specifically designed for psycholinguistic experiments. The Harvard General Inquirer is yet another psycholinguistic lexicon that contains 182 categories. The words in the lexicon are also tagged with two large valence categories, namely, negative and positive among other psycholinguistic categories, such as words of pain, pleasure, virtue and vice, words symbolising understatement, overstatement, etc. Following our earlier research [11], we also considered 14 different features from the MRC Psycholinguistic lexicon. These features were ratings of Familiarity, Concreteness, Kucera-Francis frequency, number of phoneme and syllables, Brown verbal frequency, imagability and age of acquisition [8]. Sensicon is a sensorial lexicon, containing words with sense association scores for 5 human senses which are Smell, Sight, Taste, Hearing, and Touch. Sensicon provides a numerical membership which indicates the extent to which a sense, among 5 senses that humans possess is used to discern a certain concept. We use this lexicon into both the models. We perform feature ablation using Pearson correlation of lexicon features vs Personanity and Values types. Finally, classifiers are trained using only those features which positively contribute to a particular Personality/Values type.

Non-linguistic Features: Features such as social network structures can aid find the underlying Personality and Values of an individual. Facebook network properties including

6. http://www.psych.rl.ac.uk/
8. To get these MRC features, we use the following API: http://ota.oucs.ox.ac.uk/headers/1054.xml.
betweenness, transitivity, centrality, network size and density were calculated. These values have been provided as a part of the released Facebook Personality corpus.

Some features distinctive of the Twitter Values Corpus were total number of tweets per user, number of likes per tweet, average time difference between two tweets, number of favourites and re-tweets. The centrality scores on the networks of the individuals’ followers and friends were also used.

Speech-act Features: The communication patterns of human beings – visually, verbally or via textual conversation, might be indicative enough of their Personality and Values traits. By following the hypothesis of [13], speech-act features were added to our feature list for the classification of Personalities. However, for this experiment we restrict speech-act classes into 11 major categories: Yes-No Question (YN), Action Directive (AD), Appreciation (AP), Wh Question (Wh), Statement Non-Opinion (SNO), Response Acknowledgement (RA), Yes Answers (YA), Thanking (T), Apology (A), Statement Opinion (SO) and others (O). We scraped about 7K utterances from Facebook and Quora pages, they were also annotated by us manually. Using this corpus, we build a Support Vector Machine based speech-act classifier. The features used were: top 20% most frequent word bigrams (Bag of Words), existence of question marks, existence of “wh” words, POS tags distributions, occurrence of “thanks/thanking” words and sentiment lexicons such as NRC sentiment lexicon10, SentiWordNet11 and WordNet Affect11. With 10-fold cross validation, an average F-score of 0.69 was obtained. The corpus distribution, category wise and the performance of the classifier are elucidated in Table 1.

Table 1. Speech-act class distributions in corpus along with the (a) performance and (b) error-rate of the speech-act classifier.

<table>
<thead>
<tr>
<th>Speech-act (Class distributions in %)</th>
<th>SNO</th>
<th>Wh</th>
<th>YN</th>
<th>SO</th>
<th>AD</th>
<th>YA</th>
<th>T</th>
<th>AP</th>
<th>RA</th>
<th>A</th>
<th>O</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) F-Score</td>
<td>0.45</td>
<td>0.88</td>
<td>0.88</td>
<td>0.72</td>
<td>0.45</td>
<td>0.60</td>
<td>0.72</td>
<td>0.60</td>
<td>0.12</td>
<td>0.12</td>
<td>0.69</td>
<td></td>
</tr>
<tr>
<td>(b) Error-Rate (in %)</td>
<td>4.1</td>
<td>1.1</td>
<td>1.6</td>
<td>2.1</td>
<td>3</td>
<td>1.6</td>
<td>1.2</td>
<td>3.7</td>
<td>3.9</td>
<td>2.3</td>
<td>3.6</td>
<td>2.65</td>
</tr>
</tbody>
</table>

Classifying social media conversational texts into specific speech-act classes is still an under-research problem, which is a separate research problem by its own right. We do realize that the developed speech-act classifier has not reached the de-facto status, but still the user specific speech-act distributions (in %) have been used as features for the psycholinguistic classifiers (Adding 11 new features) and we find a marked performance improvement of 6.12% (F-Score) on the Twitter Values Corpus.

Transfer learning: Another question might be raised here – is there any loss in transfer learning as many of the classifiers are applied across domains - e.g., the dialog act systems were trained on Facebook and Quora corpus, then applied on Twitter; the Personality classifier was trained on a Facebook corpus, and applied on Twitter.

For the Personality analysis there is no publicly available Twitter dataset; therefore it is difficult to assess transfer-loss for the Personality analysis. For Values model our classifier is trained and tested on Twitter corpus.

In order to examine the transfer loss for the Speech Act classifier on Twitter corpus, we have further annotated a small corpus consisting of 1,100 tweets - 100 for each class. We ran our previous speech act classifier – which was trained on Facebook and Quora data on the newly annotated corpus. We report the error rate in Table 1(b). We observe that the average error-rate is as low as 2.65% which concludes that the Speech Act Classifier can be used for the classification of tweets.

4.1 Evaluation of Personality and Values Models

Identifying the Personality and Values models can be approached both as a multi-class problem (preferred by computational people) and as a multi-valued regression problem (preferred by psychologists). Here the former view is handled by three traditional types of classifiers, namely Support Vector Machines (SVM, Linear Kernel), Multinomial Naïve Bayes (MNB), and Random Forests (RF), for which each Personality/Value type is subdivided into Yes class (positively oriented), and No class (negative oriented), resulting 10 classes ($5 \times 2$) for the Personality classifier and 20 classes ($10 \times 2$) for Values classifier. The multi-valued regression view is handled by Simple Logistic Regression (LR), using real Personality and Values scores (scaled to the interval [0,1]). We observe that our SVM-based Personality model achieves an average F-Score of 0.80, outperforming the best competing system [4] (F-Score of 0.73) as reported in Figure 2. Our Values model obtains an average F-score of 0.82. Class-wise performance of both the models is reported in Table 2.

In regression setup, the average correlation coefficient for Personality and Values models achieve average correlation (Pearson) of 0.38 and 0.29 respectively as shown in Tables 3 and 4. Respective correlation scores may be considered as moderate.

5 UNDERSTANDING THE SOCIETAL SENTIMENT

Once the automatic Personality and Values models have been produced, they can be applied to the SNAP Twitter
community dataset (see Section 3.3). To understand whether people within the same social network community conceive homogeneous characteristics w.r.t. their Personality and Values, we measure randomness for each dimension separately using Shannon’s Entropy. It is a measure of the uncertainty. Higher entropy scores indicate more diversity (low similarity). The analysis of entropy scores in each community would in turn reveal the semantic interpretation of societal sentiment.

As stated before, Personality and Values categories are interconnected to each other and influence one another, since the pursuance of any of the Personality/Values types results, either in accord with another value (e.g., conformity and tradition) or an obverse with an one other value (e.g., benevolence and achievement). We recall that both the psychology models — Personality and Values models — support fuzzy membership. Therefore, the goal is to understand if certain communities with lesser entropy scores for at least one or more Personalities/Values dimensions could be treated as homogeneous. Of course, one can argue that power oriented people can make friends or form a community with hedonic or universal people, and similarly that extroverts can handle neurotic friends. We certainly believe that such cross categorical relationships exist in real life; but here our aim is to understand whether homogeneity is perceived to a large extent within these societal communities.

For both Personality and Values models, we show the cross-entropy using heat-maps in Figures 3(a) and 3(b) respectively. In these figures, red shades represent higher entropy score, blue shades represent lower entropy scores. To understand the natural preference of people belonging to same community could be understood from these tables. For example, the first row in Figure 3(b) and Figure 3(a) represent all the communities having less entropy scores for achievement and openness respectively. Less entropy scores indicate higher similarity.

The obtained entropy scores differ greatly across communities. Therefore, the entropy scores are normalised as follows to keep the range between 0 and 1: 

\[ x_{\text{scaled}} = \frac{x - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}} \]

We further normalize the entropy scores based on the community size. Finally, communities below a experimentally chosen threshold (median of \( x_{\text{scaled}} \) Value) are considered for analysis.

There are several interesting observations in these figures. For example, in Figure 3(b), one can observe that the Security people tend to feel uncomfortable with achievement oriented group, as Security (SE) oriented people always want to be safe, and not very eager to break societal rules. Achievement (AC) oriented people, on the other hand, are always very eager to achieve their personal ambitions, and are ready to take any risks to achieve their goals [14]. Another interesting observation is that Traditional (TR) people always find difficulties to stay inside other types of communities. The reverse is also true – other people also find difficulties to join TR oriented groups (row TR), resulting in very low entropy scores for the whole row. Similar trends can be seen for power (PO) groups, and for the conscientiousness (C) vs extroversion (E) personalities in Figure 3(a) [14]. On the other hand, self-direction (SD) people find it very hard to fit into a conformity (CO) group as SD (Self-Direction) oriented people want to lead life on their own rules. A few scaling factors are also interesting. For example, in a group of agreeable personalities (row 4 in Figure 3(a)), open people (col. 1) find themselves quite comfortable, whereas the reverse is not true.

Comparison with Random Users: One might further ask whether a population drawn randomly possesses the similar psycho-sociological properties as that of the naturally formed population (such as in the SNAP dataset). Reports of psycho-sociological analysis always depends on how the target population is chosen. To understand this notion empirically, we perform a random sample test. For the experiment we chose 5,000 Twitter users and created a random set. This experiment was repeated 20 times. The psycho-sociological traits of those random samples with users present in the SNAP dataset is reported in Figure 4 (for the sake of brevity we only report curves corresponding to 5 random samples along with the SNAP dataset). We observe that the obtained Personality and Values distributions in each set largely differ from each other. Moreover, the difference is statistically significant according to the t-test with 95% confidence interval.

6 Community Detection using Personality and Values

So far, we have analyzed Values and Personality of individuals belonging to different network communities. We hypothesize that these two features can be used as node-centric attributes to detect better community structure in the

Fig. 2. Performance of Personality and Values Models with different feature sets added incrementally.

Fig. 3. Heatmap representing cross-relational entropy scores among different (a) Personality vs Personality and (b) Values vs Values.

Both the algorithms allows us to control the weights of network and node features. We consider the Twitter network and three sets of features – the network information, the Personality and Values features. We run them with Values and Personality feature sets separately along with network information. We use real values obtained from the regression models directly into the feature vector. We report the accuracy of CESNA and T-CME for individual feature settings separately in Table 5. Finally, we combine all these features together with equal weights given to node attributes and network information and report the accuracy in Table 5.

Community Evaluation Metrics. Since we know the ground-truth community structure of the Twitter network as discussed in Section 3.3, we compare the detected community structure with the ground-truth using the following standard metrics – Normalized Mutual Information (NMI), Adjusted Rand Index (ARI), Purity (PU). The value of the first three metrics ranges between $(-1)$ and $1$. The more the value of the metrics, the better the detected community structure resembles with the ground-truth.

Results. Table 5 presents the accuracy of both the algorithms for different feature sets. Overall, we observe that considering the node attributes along with the network features always improves the performance of the algorithms as opposed to considering only the network information. Considering all the features, CESNA (T-CME) achieves $7\%$ (8.6%), $11.41\%$ (11.11%) and $9.23\%$ (7.35%) higher NMI, ARI and PU respectively compared to the case when only network features are used. This result corroborates with our hypothesis that addition of node attributes indeed enhances the performance of the community detection algorithms.

### Table 5. The accuracy of T-CME and CESNA (within parenthesis) with different feature sets.

<table>
<thead>
<tr>
<th>SL #</th>
<th>Used features</th>
<th>NMI</th>
<th>ARI</th>
<th>PU</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a)</td>
<td>Network information</td>
<td>0.57 (0.38)</td>
<td>0.61 (0.63)</td>
<td>0.66 (0.68)</td>
</tr>
<tr>
<td>(b)</td>
<td>(a) + Value feature</td>
<td>0.57 (0.61)</td>
<td>0.61 (0.62)</td>
<td>0.66 (0.68)</td>
</tr>
<tr>
<td>(c)</td>
<td>(b) + Personality feature</td>
<td>0.59 (0.60)</td>
<td>0.64 (0.65)</td>
<td>0.69 (0.70)</td>
</tr>
<tr>
<td>(d)</td>
<td>All</td>
<td>0.61 (0.63)</td>
<td>0.68 (0.70)</td>
<td>0.71 (0.73)</td>
</tr>
</tbody>
</table>

7 Conclusion & Future Directions

In this paper, we have attempted to establish a correlation between the Personality and Values of individuals belonging to the same community in social networks. In particular, the contributions of our paper are fivefold: (i) development of Schwartz’ Values Twitter corpus, which we will release soon for research purpose; (ii) design of automatic systems to categorize users based on their Values; (iii) development of advanced models to identify users’ personality traits; (iv) understand the semantic interpretation of social communities, which to the best of our knowledge is addressed for the first time in this paper, and (v) augmentation of Personality and Values of individuals as node attributes along with the network information in order to detect better community structure. However, the terse nature of social media text is a problem for any system development. In future, we would like to study the demographic psycholinguistic variations across social network communities. Moreover, we are keen to apply the psycholinguistic models for other applications such as Internet advertising, surveillance over social media for counter-terrorism, etc.

References


Tushar Maheshwari is an undergraduate student at Department of Computer Science and Engineering of the Indian Institute of Information Technology, Sri City, India.
Aishwarya N. Reganti is an undergraduate student at Department of Computer Science and Engineering of the Indian Institute of Information Technology, Sri City, India.

Upendra Kumar is an undergraduate student at Department of Computer Science and Engineering of the Indian Institute of Information Technology, Sri City, India.

Tanmoy Chakraborty is an Assistant Professor at Department of Computer Science and Engineering, Indraprastha Institute of Information Technology Delhi, India. His broad research interests are Social Network Analysis, Data Mining and NLP.

Amitava Das is an Assistant Professor at Department of Computer Science and Engineering of the Indian Institute of Information Technology, Sri City, India. His research interests broadly span three areas and more specifically their intersection: Natural Language Processing, Social Media Analysis, and Computational Creativity.