A Societal Sentiment Analysis: Predicting the Values and Ethics of Individuals by Analysing Social Media Content

Aishwarya Reganti†, Tushar Maheshwari†, Upendra Kumar†, Tanmoy Chakraborty*, and Amitava Das†
†Indian Institute of Information Technology, Sri City, AP, India
*University of Maryland, Maryland, USA
†{aishwarya.r14, tushar.m14, upendra.k14,amitava.das}@iiits.in,
*tanchak@umiacs.umd.edu

ABSTRACT

In network science, a community is considered to be a group of nodes densely connected internally and sparsely connected externally. Communities are imminent in social media networks where every user has a list of acquaintances who are connected to another set of users, and many a times these users are mutually connected to each other. A big closed unit of such users can be rightly called a community or circle in a social media network. Detecting and analyzing communities from social networks has attracted immense attention over the last decade. However, the semantic interpretation of a community is hardly studied. In this paper, we attempt to understand *whether individuals in a community possess similar personalities, values and ethical background*. To this end, we collect Twitter values corpus, extract the network communities and propose automatic models to determine personality, values and ethics of individuals. Various experiments are performed to understand the characteristics or blend of characteristics of individuals within a community.

1. INTRODUCTION

Detecting and analyzing dense groups or communities from social and information networks has attracted immense attention over the last decade [4]. Since community detection is an *ill-defined problem*, several heuristics were proposed to detect communities from the network structure [3]. However, the semantic interpretation of a community, i.e., the behavior of individuals in a community has hardly been addressed. This paper presents psycholinguistic study in order to understand the behavior of individuals forming communities in social networks. We use two psycholinguistic models to identify the behavior of individuals thoroughly. The Personality model is used to understand the characteristics or blend of characteristics at individual level, whereas the values and ethics models are used to understand and analyze interpersonal dynamics of societal sentiment.

The Big 5 personality traits [5], aka the five factor model (FFM), is the mostly used *personality model*. The five factors are **Openness (O)**: A personality trait possessed by individuals who are imaginative and 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 five model is also represented using the acronym OCEAN.

In order to define *societal sentiment*, we utilize the well-established "Schwartz Theory of Basic Human Values" [11], which defines ten basic and distinct personal values, namely: **Achievement (AC)**: The value here comes from setting goals and then achieving 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.

2. CORPUS

To start with, we asked a very fundamental question - *whether social media is a good proxy of the original society or not*. We grounded our corpus collection based on the conclusion drawn by [2] that people, in general, do not use virtually desired/bluffed social media profiles to promote an idealized-virtual-identity.

2.1 Twitter Values Corpus

The standard method for any psychological data collection is through self-assessment tests, popularly known as psychometric tests. Self-assessments were obtained using a fifty-item long gender specific (male/female) version of the Portrait Values Questionnaire (PVQ) (Schwartz et al.,2001). 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) answer the PVQ questions honestly and (ii) provide their Twitter ids so that their tweets could be crawled. However, we faced several challenges while working with Twitter and therefore a number of iterations, human interventions and personal commu-

ACM ISBN 978-1-4503-2138-9.
DOI: 10.1145/1235

nications had to be done to resolve all these issues. In the end, 367 unique user's data were gathered containing PVQ answers along with user's tweets. The highest number of tweets for a user was quite high (15K) and the lowest number of tweet per user was a only 100 (average is 1,608). We also ensured that participants are native English speakers from various cultures and ethnic backgrounds.

2.2 Facebook Personality Corpus

In the recent years, there have been a lot of research activities on automated identification of various personality traits of an individual from their language usage and behaviour in social media. One milestone in this area is the "Workshop and Shared Task on Computational Personality Recognition" (WCPR) in 2013. They released a Facebook corpus, which consists of 10,000 Facebook status updates of 250 users and their Facebook network properties, labeled with personality traits. Among the 8 teams participated in this shared task, the best performance was achieved by [14] with an average F-Score of 0.72. In Section 3, we will show that our model outperforms their system by achieving an average F-Score of 0.78.

2.3 Twitter Community Corpus

We collected the Twitter network, released by SNAP [6] (nodes: 81,306, edges: 1,768,149). The users are distinguished by their Twitter id's. This dataset was widely used in the study of community detection [8, 9]. Several previous studies [16, 12, 17] showed that it contains a strong community structure. We further enriched the dataset by crawling the tweets of each user, required for our personality and values models.

To this end, we use Personality and Values corpora for designing our models and use them to analyze the community structure present in the SNAP dataset.

3. COMPUTATIONAL PERSONALITY AND VALUES MODELS

The state-of-the-art sentiment analysis systems analyze any fragment of text in isolation. However, in order to design Big five personality model and Schwartz model classifier, psycholinguistic analysis is required.

Here, we discuss various features and methods used for the automatic personality and values identification. Our models are inspired by the research reports published in the Workshop and Shared Task on Computational Personality Recognition [7]. We experiment with several machine Learning algorithms such as Support Vector Machine (SVM), Multinomial Naive Bayes (mNM), Simple Logistic Regression (LR), and Random Forest (RF). Among them SVM (with linear kernel) turns out to be the best method.

Linguistic Features: We use three different Psycholinguistic lexicon features: Linguistic Inquiry Word Count (LIWC) [10], Harvard General Inquirer [13] and MRC psycholinguistic database [15]. LIWC is a well developed hand-crafted lexicon, containing 69 different categorical words, specifically designed for psycholinguistic experiments. Harvard General Inquirer lexicon contains 182 categories which include two large valence categories – positive and negative, and other psycholinguistic categories such as words of pleasure, pain, virtue and vice, words indicating overstatement, understatement etc. We include 14 different features from the MRC Psycholinguistic lexicon namely, number of phoneme & syllables, Kucera-Francis frequency, Kucera-Francis number of categories, Kucera-Francis number of samples, Thorndike-Lorge frequency, Brown verbal frequency, ratings of Familiarity, Concreteness, imagability & age of acquisition[1].

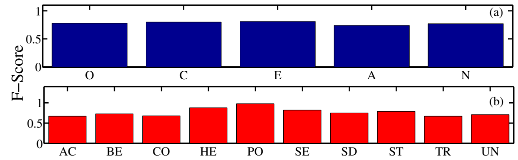

Figure 1: Performance of (a) Personality (b) Values models.

We perform feature ablation and also analyze (Pearson) correlations of each lexicon features vs personality and value types. Finally classifiers are trained using only the features which are contributing for a particular personality/value type.

Non-Linguistic Features: Social network structure is very useful to predict any individual's intrinsic personality and value. Facebook network properties including network size, betweenness centrality, density and transitivity, are provided as a part of the released Facebook personality corpus [7]. For the Twitter values corpus, total number of tweets/or messages of one user, total number of likes, average time difference between two tweets/messages, total number of favourites and re-tweets on all the tweets/messages by one user and their in-degree and out-degree centrality scores on network of friends and followers are used as features. The degree centrality is calculated for each vertex v in a given graph *G(V,E)* with $|V|$ vertices and $|E|$ edges, and is defined as: $C_D = deg(v)$.

The Personality and Values models achieve average F-scores of 0.78 and 0.70 respectively. Class wise F-Scores for both the systems are reported in the Figure 1.

4. UNDERSTANDING THE SOCIETAL SENTIMENT

In order to analyse whether people within the same community circle have some homogeneity with each other's personality and background values/ethics, Shannon's Entropy [1] is calculated for each dimension. This is a measure of the uncertainty in the probability distribution. Then we analyze patterns of those communities, which reveal the semantic interpretation of societal sentiment.

Personality and values categories are interconnected and influence each other, since the pursuit of any of the personality/value types results, either in accordance with one another (conformity and security) or a conflict with at least another value (benevolence and power). It must be noted that both the psychology models – Personality and Values models – support fuzzy membership. Therefore, our 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. However, one can also argue that power oriented people can also make friends or form a community with hedonic or universal people, and similarly extroverts can really handle neurotic friends. We also believe that such cross categorical relationship in real life certainly exists, but here our aim is to understand whether homogeneity is perceived to a large extent in these societal communities.

Cross-relational entropies both for Personality and Values models are reported in Table 1 and Table 2 respectively. In these tables, rows represent communities having less entropy scores for the corresponding psychological dimension. For example in the Table 1 the first row AC (Achievement) represents all the communities having less entropy scores for achievement. Columns represent the fuzzy orientations of the community members in rest of the dimensions. Resulting entropy scores for different personality and values differ greatly for various communities. Therefore we normalize entropy scores using the formulation: $x_{scaled} = \frac{(x - x_{min})}{x_{max} - x_{min}}$ which keeps the value distribution between 0 and 1. We then choose com-

[1]To get these MRC features, we used the following API: http://ota.oucs.ox.ac.uk/headers/1054.xml.

munities below the calculated threshold (mid of x_{scaled} value).

In Table 1, one can observe from the column-wise distribution of the AC (row 1) that the SE (Security) (col. 6) people find it difficult to manage in any achievement oriented group, as SE people always want to be safe and are unwiling to go against rules, whereas AC people are always very keen on achieving their goals and are ready to take risks for the same. Another interesting observation is that traditional (TR) (col. 9) people can hardly manage themselves in any other oriented group. Similarly in traditional (TR) (row 9) oriented groups other people hardly join, resulting very low entropy in almost all the entries corresponding to TR row. Similar trend can be seen for power (PO) (row 5) groups in Table 1 and for the conscientiousness (C) (row 2) vs extroversion (E) (col. 3) personalities in Table 2. On the other hand, self-direction (SD) (col. 7) people find it very hard to fit into a conformity (CO) (row 3) group as SD oriented people want to run their lives by their own rules. A few scaling factors are also interesting. For example, in a group of agreeable personalities (row 4) an open people (col. 1) find themselves quite comfortable, whereas the reverse is not true. e in a corpus of size 367 users data. Therefore, we need more data for two obvious reasons – (i) to train the system for better performance, and (ii) for more precise analysis of the communities. We are in the process of collecting more data.

Class	AC	BE	CO	HE	PO	SE	SD	ST	TR	UN
AC	–	0.01	0.13	0.09	**0.76**	**0.78**	0.08	0.78	**0.79**	0.17
BE	0.01	–	0.13	0.18	**0.79**	0.18	0.18	0.18	**0.83**	0.27
CO	0.00	0.01	–	0.09	**0.76**	0.09	**0.89**	0.89	**0.76**	0.14
HE	0.01	0.00	0.15	–	**0.73**	0.00	0.01	0.00	**0.78**	0.09
PO	0.01	0.00	0.13	0.06	–	0.07	0.06	0.07	**0.79**	0.15
SE	0.00	0.00	0.15	0.00	**0.73**	–	0.02	0.02	**0.77**	0.09
SD	0.00	0.00	0.16	0.00	**0.72**	0.01	–	0.00	**0.77**	0.10
ST	0.01	0.00	0.14	0.01	**0.73**	0.01	0.00	–	**0.79**	0.09
TR	0.00	0.00	0.10	0.05	**0.75**	0.05	0.05	0.06	–	0.13
UN	0.01	0.00	0.12	0.00	**0.73**	0.00	0.01	0.01	**0.75**	–

Table 1: Values and Ethics model

Class	O	C	E	A	N
O	–	**0.76**	**0.74**	0.31	0.32
C	0.00	–	**0.56**	0.21	0.20
E	0.01	**0.59**	–	0.21	0.19
A	0.00	**0.57**	**0.55**	–	0.18
N	0.20	**0.59**	**0.54**	0.19	–

Table 2: Personality model

Table 1 & 2: Cross-relational entropy for Personality and Values

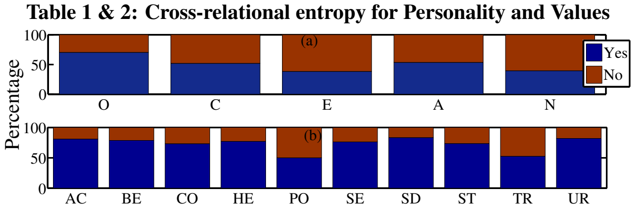

Figure 2: Class distribution in (a) Personality and (b) Values corpora.

It is prevalent from the results that the values of Shannon's Entropy would be a maximum of 1 since there are two classes. In some of the circles we get the value as 1 which clearly indicates that there is no similarity however, for a major number of circles we found considerable good average for values and ethics and personality which was below 0.2 and 0.4 for both the respective classes. Hence, it can be said that the community network structure has a subtle relation with personality and Values and Ethics. Moreover, Table 1 2 illustrate the fuzzy relationship between the psycholinguistic models.

5. CONCLUSION & FUTURE DIRECTIONS

In this paper, we have attempted to establish a correlation between the personalities and values of individuals belonging to the same community. In particular, the contributions of our paper are fourfold: (i) the 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 ethical beliefs; (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. Our next endeavor would be to design more accurate systems and examine whether Personality and Values models could be used as features along with network features for community detection.

6. REFERENCES

[1] J. AczÃl'l and Z. DarÃşezy. *On Measures of Information and Their Characterizations*. Mathematics in Science and Engineering. Elsevier Science, 1975.

[2] M. D. Back, J. M. Stopfer, S. Vazire, S. Gaddis, S. Schmukle, B. Egloff, and S. D. Gosling. Facebook profiles reflect actual personality, not self-idealization. *Psychological Science*, 21:372–374, 2010.

[3] T. Chakraborty, A. Dalmia, A. Mukherjee, and N. Ganguly. Metrics for community analysis: A survey. *CoRR*, abs/1604.03512, 2016.

[4] S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75 – 174, 2010.

[5] L. R. Goldberg. An alternative" description of personality": the big-five factor structure. *Journal of personality and social psychology*, 59(6):1216, 1990.

[6] J. Leskovec and A. Krevl. SNAP Datasets: Stanford large network dataset collection.

[7] D. Markovikj, S. Gievska, M. Kosinski, and D. Stillwell. Mining facebook data for predictive personality modeling. In *ICWSM*, 2013.

[8] J. Mcauley and J. Leskovec. Discovering social circles in ego networks. *ACM Trans. Knowl. Discov. Data*, 8(1):4:1–4:28, Feb. 2014.

[9] T. P. Peixoto. Model selection and hypothesis testing for large-scale network models with overlapping groups. *Phys. Rev. X*, 5:011033, Mar 2015.

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

[11] S. H. Schwartz. Universals in the content and structure of values: theoretical advances and empirical tests in 20 countries. pages 1–65. In Advances in experimental social psychology (Vol. 25), 1992.

[12] J. Shang, L. Liu, X. Li, F. Xie, and C. Wu. Epidemic spreading on complex networks with overlapping and non-overlapping community structure. *Physica A: Statistical Mechanics and its Applications*, 419(C):171–182, 2015.

[13] P. J. Stone, D. C. Dunphy, and M. S. Smith. The general inquirer: A computer approach to content analysis. 1966.

[14] B. Verhoeven, W. Daelemans, and T. De Smedt. Ensemble methods for personality recognition. *ICWSM-13*.

[15] M. Wilson. Mrc psycholinguistic database: Machine-usable dictionary, version 2.00. *Behavior Research Methods, Instruments, & Computers*, 20(1):6–10, 1988.

[16] J. Yang and J. Leskovec. Overlapping community detection at scale: A nonnegative matrix factorization approach. In *WSDM '13*, pages 587–596.

[17] Y. Yang, C. Lan, X. Li, B. Luo, and J. Huan. Automatic social circle detection using multi-view clustering. In *CIKM '14*.