Cosmopolitan Mumbai, Orthodox Delhi, Techcity Bangalore: Understanding City Specific Societal Sentiment

Aishwarya N Reganti, Tushar Maheshwari, Upendra Kumar, Amitava Das

Indian Institute of Information Technology, Sri City, Andhra Pradesh, India
{aishwarya.r14, tushar.m14, upendra.k14, amitava.das}@iiits.in

Abstract

The paper reports work on investigating societal sentiment using the Schwartz values and ethics model, and applying it to social media text of users from 20 most populous cities of India to represent geo-specific societal sentiment map of India. For the automatic detection of societal sentiment we propose psycholinguistic analysis, that reveals how a user's social media behaviour and language is related to his/her ethical practices. India is a multi-cultural country, values and ethics of each Indian are highly diverse and dependent on the region or society s/he belongs to. Several experiments were carried out incorporating Linguistic Inquiry Word Count analysis, n-grams, topic modeling, psycholinguistic lexica, speech-acts, and non-linguistic features, while experimenting with a range of machine learning algorithms including Support Vector Machines, Logistic Regression, and Random Forests to identify the best linguistic and non-linguistic features for automatic classification of values and ethics.

1 Introduction

Indian sub-continent is known for its diversity and multiplicity, with 29 states and 7 union territories each region has a culture of its own. India's plethora of rich cultural diversity is reflected in the fact that each region is unique in terms of the values and ethics borne by the individuals belonging to these regions or society. For example, Punjabis are fun loving and energetic, South Indians are reserved and introverts as compared to North Indians, Mumbaikars and Delhiites are outgoing and socially confident etc. The main objective of this paper is to demonstrate that these regional ethics can be inferred from social media languages and behavior using statistical analysis.

In the last decade there have been significant efforts in opinion and sentiment mining as well as inferring emotion from text. Classical sentiment/emotion analysis systems classify text into either one of the classes positive, negative or neutral, or into Ekman's classes of happy, sad, anger, fear, surprise, and disgust. Personality models [John and Srivastava1999] can be seen as an augmentation to the basic definition of sentiment analysis, where the target is to understand sentiment/personality at person level rather than only at message level. Here in this paper our motivation is to augment to one more level i.e., to understand societal sentiment, i.e., values and ethics. To understand the societal sentiment we borrow the psychological values and ethics model introduced by schwartz1990toward, defines ten basic and distinct ethical values (henceforth only values). The definitions of these 10 values are as following. ***Achievement***: sets goals and achieves them; ***Benevolence***: seeks to help others and provide general welfare; ***Conformity***: obeys clear rules, laws and structures; ***Hedonism***: seeks pleasure and enjoyment; ***Power***: controls and dominates others, controls resources; ***Security***: seeks health and safety; ***Self-direction***: wants to be free and independent; ***Stimulation***: seeks excitement and thrill; ***Tradition***: does things blindly because they are customary; ***Universalism***: seeks peace, social justice and tolerance for all [Schwartz2012]. All these basic ten values can be further grouped into higher order (super class: SC) groups which respectively are: **Self-Enhancement** (SC1) {Achievement, Power, Hedonism}, **Openness to change** (SC2) {Stimulation, Self-directions, Hedonism}, **Self-Transcendence** (SC3) {Universalism, Benevolence}, **Conservation** (SC4) {Security, Tradition,

D S Sharma, R Sangal and A K Singh. Proc. of the 13th Intl. Conf. on Natural Language Processing, pages 167–176, Varanasi, India. December 2016. ©2016 NLP Association of India (NLPAI)

Conformity}. In addition to identifying the ten basic values, Schwartz' theory also explains how the values are interconnected and influenced with each other, since the pursuit of any of the values results in either an accordance with one another (e.g., Conformity and Security) or a conflict with at least one other value (e.g., Benevolence and Power).
These ten basic values are related to various outcomes and effects of a person's role in a society [Argandoña2003, Agle and Caldwell1999, Hofstede et al.2010, Rokeach1973]. The values have also proved to provide an important and powerful explanation of consumer behaviour and how they influence it [Kahle et al.1986, Clawson and Vinson1978]. Moreover, there are results that indicate how values of workforce and ethical practices in organizations are directly related to transformational and transactional leadership [Hood2003]. We believe that these kind of models may become extremely useful in the future for various purposes like Internet advertising (specifically social media advertising), community detection, computational psychology, recommendation systems, sociological analysis (for example East vs West cultural analysis) over social media. Customized products can be crafted for specific regions of the country by analysing the collective values of that region. In order to experiment, a twitter corpus was collected and annotated with Schwartz values. A range of machine learning techniques were then utilized to classify an individual's ethical practices into Schwartz' classes by analyzing the user's language usage and behaviour in social media. The borders between the motivators are artificial and one value flows into another. Such overlapping and fuzzy borders between values make the computational classification problem more challenging.
The paper is organized in sections with Section 2 introducing related work in this area. The details of the corpora collection and annotation are given in Section 3. In Section 4 various experiments on automatic value detection are reported, Section 5 illustrates the results obtained for Indian cities and inferences that can be drawn while Section 6 discusses the performance of the psycholinguistic experiments and mentions possible future directions.

2 Related Works

The present day state-of-the-art sentiment analysis systems look at a fragment of text in isolation. Therefore textual features and models proposed and discussed for such purposes is quite different than our current research need. Henceforth, we are focusing our discussion on the previous research on automatic personality analysis which is more closer to our research.

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 in the year of 2013. Two corpora were released for this task. One was the Facebook corpus, which consisted of about 10,000 Facebook status updates of 250 users, plus their Facebook network properties, labelled with personality traits. markovikj2013mining and verhoeven2013ensemble achieved encouraging results on personality classification task. The following two paragraphs consolidate our discussion on various features and methods used by researchers in the workshop. Two kinds of features were put into use: linguistic and non-linguistic.

Linguistic Features: The participating teams tested several linguistic features. As a general fact, it is known that n-grams are the most famous and useful features for any kind of textual classification. Therefore, teams tested various features like unigrams, bigrams, and trigrams. We also noticed that Linguistic Inquiry Word Count (LIWC) features were used by all the teams as their baselines. LIWC [Pennebaker et al.2001] is a well developed hand-crafted lexicon. It has 69 different categorical words, specifically designed for psycholinguistic experiments. Another psycholinguistic lexicon called MRC [Wilson1988] was also used by a few teams. Lexicons like SentiWordNet [Baccianella et al.2010], WordNet Affect strapparava:369:2004:lrec2004, categorical features like part-of-speech (POS), a few other word level features like capital letters, repeated words were also used. Two more important textual features were discussed by the participating teams. Linguistic nuances introduced by tomlinson2013predicting, is the depth of the verbs in wordnet troponymy hierarchy. Speech act features have been discussed by appling2013towards. Authors manually annotated the given Facebook corpus with speech acts and reported their correlation with personality traits of the respective users.

Non-Linguistic Features: Facebook network properties including network size, betweenness centrality, density and transitivity, provided as a part of the released dataset, were used by all the teams.

Drawing inspiration from the learning of the above mentioned research, we have experimented with all the above mentioned linguistic and non-linguistic features and proposed a few new features. Effectiveness and performance impact of each feature on the classification result has been detailed and discussed in the section 4.

3 Corpus Acquisition

At the very beginning of this research, we asked ourselves very fundamental question - *if social media is a good proxy of the original society?* backetal2010 and GolbeckEA:11 provide empirical answers to this question. Their results respectively indicate that, in general, people do not use virtual desired/bluffed social media profiles to promote an idealized-virtual-identity and that a user's personality can be predicted from his/her social media profile. This does not mean that there are no outliers, but we grounded our corpus collection on the assumption that is true for a major portion of the population.

A standard method of psychological data collection is through self-assessment tests which are popularly known as psychometric tests. Self-assessments were obtained using a 50-question version of the Portrait Values Questionnaire (PVQ) [Schwartz et al.2001]. This questionnaire consists of questions where the participant's answer must be a score rating from 1-6 Likert scale. A rating of 1 means "not like me at all" which progressively changes to 6 meaning "very much like me". An example question is "*He likes to take risks. He is always looking for adventures.*" where the user should answer while putting himself in the shoes of "He" in the question. Once all the questions in the PVQ have been answered, a value score is generated using an averaging formula for each value consisting of scores of the related questions.

3.1 Twitter Values Corpus

In the first quarter of 2016, the micro blogging service Twitter averaged 310 million monthly active users,[1] with around 6,000 tweets being posted every second. Unlike other popular social media like Facebook, Twitter provides open access to its data. Therefore, Twitter came as the natural choice for the data collection. The data collection was crowd-sourced using Amazon Mechanical Turk as a service, while ensuring that the participants came from various cultures and ethnic backgrounds: the participants were equally distributed, and consisted of Americans (Caucasian, Latino, African-American), Indians (East, West, North, South), and a few East-Asians (Singaporeans, Malaysian, Japanese, Chinese). The selected Asians were checked to be mostly English speaking.

The participants were requested to answer the 50-question version of PVQ and to provide their Twitter IDs, so that their tweets could be crawled. However, several challenges have to be addressed when working with Twitter, and a number of iterations, human interventions and personal communications were needed to resolve all the issues. For example, several users had protected Twitter accounts, so that their tweets were not accessible when using the Twitter API. In addition, many users had to be discarded since they had published less than 100 tweets, making them uninteresting for statistical analysis. The open source free Twitter API: Twitter4J[2] also has a limit of accessing only the current 3,200 tweets from any user. To resolve this issue, an open source Java application [Henrique2015] was used. At the end of the data collection process, data from 367 unique users had been gathered. The highest number of tweets for one user was 15K, while the lowest number of tweets for a user was a mere 100; the average number of messages per user in the Twitter corpus was found to be 1,608.

3.2 Geo-Specific Data for Indian Values Map

We started our data collection with a list of India's top 20 populous cities spanning almost the whole country i.e. North, South, East and West India. Names of those cities could be seen in the Table 3. Twitter GET geo/search API [Yamamoto2014] provides a facility of searching users based on latitude-longitude. We searched city specific users

[1] http://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/

[2] http://twitter4j.org/

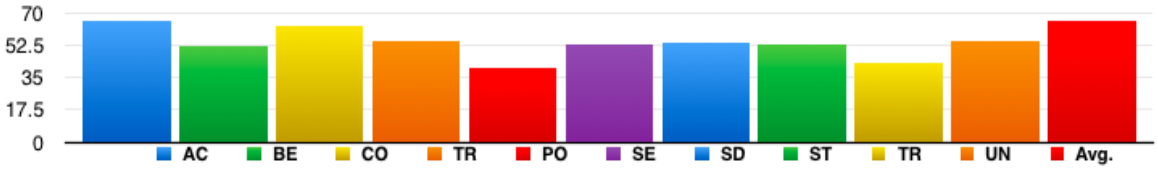


Figure 1: Values Ethics Class Distribution

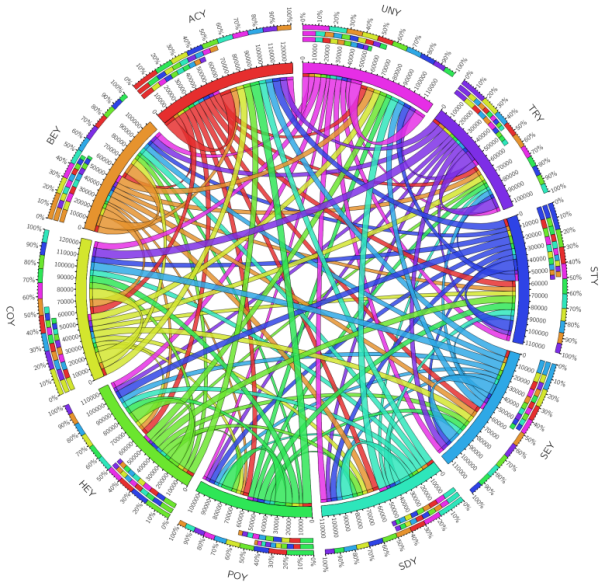


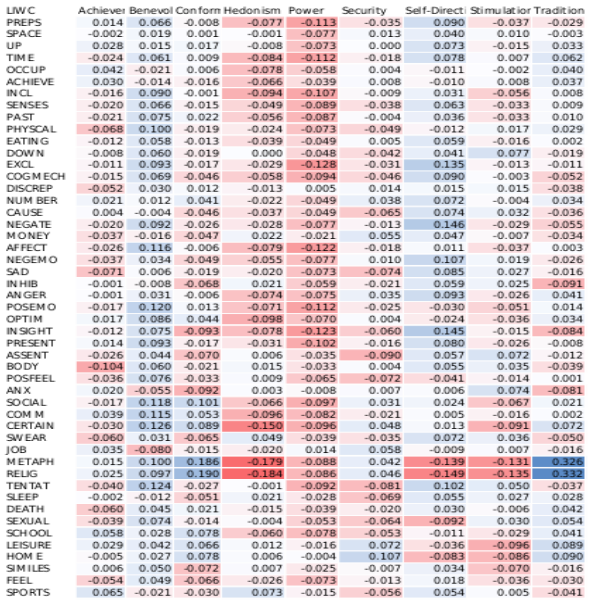
Figure 2: Values Ethics Fuzzy Class Distribution

within latitude-longitude range of 0.5°*0.5° (i.e. approximately area of 55*55 sq.km).

3.3 Post-Processing and Corpus Statistics

The scores obtained from the questionnaire were in the range 1–6 and the corpus contain crowd sourced data, so we had no control over value type distributions. For example, the Schwartz value scores for the corpus were in the following ranges: Achievement [-4.12, 3.36], Benevolence [-1.56, 3.39], Conformity [-3.35, 3.01], Hedonism [-5.18, 4.35], Power [-6.12, 2.27], Security [-2.60, 2.40], Self-Direction [-1.61, 3.40], Stimulation [-5.0, 2.63], Tradition [-4.49, 3.35], Universalism [-3.33, 3.30]. The distribution of a particular value type over a corpus was analysed using the Bienaymé-Chebyshev Inequality [Bienaymé1853, Tchébichef1867], showing that, for example, most of the Achievement instances (89%) were in the range [-2.96, 2.84]. Therefore normalization was applied to obtain equilibrium in the obtained value ranges. For all the corpora, the range of each value type was rescaled to the range of [-1,1] using the formula: $\{x_{scaled} = \frac{2*(x-x_{min})}{x_{max}-x_{min}} - 1\}$. A 1 ('Yes') or 0 ('No') class was assigned to each Schwartz value. If the score was less than 0, the class was considered to be negative, indicating absence of that value trait for the particular user, while scores ≥ 0 were considered to belong to the 'Yes' class, indicating the presence of that trait for the user.

Schwartz' model defines fuzzy membership, which means that an power-oriented individual can be achievement oriented as well, i.e., the values can overlap with each other. To understand this notion vividly, we have reported the fuzzy membership in the Figure 2 where it can be observed that the emerging bands signifies the fuzzy membership of other 9 values into one of the remaining values. For example, on careful investigation of Figure 2, we can observe that ACY oriented people's fuzzy membership is represented by outgoing red bands. The width of the bands represent the degree of membership ACY oriented people have in other values. Similarly, we can also observe that there are 10 incoming bands of 10 different colours towards ACY, which shows membership of each of the classes in ACY class.In each class, it can be observed that there is an self-arc which represents membership of each class with itself (i.e 100%). The self-arc can be used as reference in order to measure the degree of membership of each of the classes with other classes. The intricate structure of Circos (by Krzywinski18062009) figure rightly signifies how values are strongly connected with each other at societal level. Some of the interesting trends which can be observed from the Figure 2 are : power (POY) oriented people have highest overlap with achievement (ACY) oriented people, tradition oriented people have highest overlap with people high in conformity and hedonic (HEY) people have significant overlap with people seeking stimulation (STY).



LIWC	Achieve	Benevol	Conform	Hedonism	Power	Security	Self-Direct	Stimulatior	Tradition
PREPS	0.014	0.066	-0.008	-0.077	-0.113	-0.035	0.090	-0.037	-0.029
SPACE	-0.002	0.019	0.001	-0.001	-0.077	0.013	0.040	0.010	-0.003
UP	0.028	0.015	0.017	-0.008	-0.073	0.000	0.073	-0.015	0.033
TIME	-0.024	0.061	0.009	-0.084	-0.112	-0.018	0.078	0.007	0.062
OCCUP	0.042	-0.021	0.006	-0.078	-0.058	0.004	-0.011	-0.002	0.040
ACHIEVE	0.030	-0.014	-0.016	-0.066	-0.039	0.008	-0.010	0.008	0.037
INCL	-0.016	0.090	-0.001	-0.094	-0.107	-0.009	0.031	-0.056	0.008
SENSES	-0.020	0.066	-0.015	-0.049	-0.089	-0.038	0.063	-0.033	0.009
PAST	-0.021	0.075	0.022	-0.056	-0.087	-0.004	0.036	-0.033	0.010
PHYSCAL	-0.068	0.100	-0.019	-0.024	-0.073	-0.049	-0.012	0.017	0.029
EATING	-0.012	0.058	-0.013	-0.039	-0.049	0.005	0.059	-0.016	0.002
DOWN	-0.008	0.060	-0.019	0.000	-0.048	-0.042	0.041	0.077	-0.019
EXCL	-0.011	0.093	-0.017	-0.029	-0.128	-0.031	0.135	-0.013	-0.011
COGMECH	-0.015	0.069	-0.046	-0.058	-0.094	-0.046	0.090	-0.003	-0.052
DISCREP	-0.052	0.030	0.012	-0.013	0.005	0.014	0.015	0.015	-0.038
NUMBER	0.021	0.012	0.041	-0.022	-0.049	0.038	0.072	-0.004	0.034
CAUSE	0.004	-0.004	-0.046	-0.037	-0.049	-0.065	0.074	0.032	-0.036
NEGATE	-0.020	0.092	-0.026	-0.028	-0.077	-0.013	0.146	-0.029	-0.055
MONEY	-0.037	-0.016	-0.047	0.022	-0.021	0.055	0.047	-0.007	-0.034
AFFECT	-0.026	0.116	-0.006	-0.079	-0.122	-0.018	0.011	-0.037	0.003
NEGEMO	-0.037	0.034	-0.049	-0.055	-0.077	0.010	0.107	0.019	-0.026
SAD	-0.071	0.006	-0.019	-0.020	-0.073	-0.074	0.085	0.027	-0.016
INHIB	-0.001	-0.008	-0.068	0.021	-0.059	-0.021	0.059	0.025	-0.091
ANGER	-0.001	0.031	-0.006	-0.074	-0.075	0.035	0.093	-0.026	0.041
POSEMO	-0.017	0.120	-0.013	-0.071	-0.112	-0.025	0.030	-0.051	0.014
OPTIM	0.017	0.086	0.044	-0.098	-0.070	0.004	-0.024	-0.036	0.034
INSIGHT	-0.012	0.075	-0.093	-0.078	-0.123	-0.060	0.145	-0.015	-0.084
PRESENT	0.014	0.093	-0.017	-0.031	-0.102	-0.016	0.080	-0.026	-0.008
ASSENT	-0.026	0.044	-0.070	0.006	-0.035	-0.090	0.057	0.072	-0.012
BODY	-0.104	0.060	-0.021	0.015	-0.033	0.004	0.055	0.035	-0.039
POSFEEL	-0.036	0.076	-0.033	0.009	-0.065	-0.072	-0.041	-0.014	0.001
ANX	0.020	-0.055	-0.092	0.003	-0.008	0.007	0.006	0.074	-0.081
SOCIAL	-0.017	0.118	0.101	-0.066	-0.097	0.031	0.024	-0.067	0.021
COMM	0.039	0.115	0.053	-0.096	-0.082	-0.021	0.005	-0.016	0.002
CERTAIN	-0.030	0.126	0.089	-0.150	-0.096	0.048	0.013	-0.091	0.072
SWEAR	-0.060	0.031	-0.065	0.049	-0.039	-0.035	0.072	0.036	-0.050
JOB	0.035	-0.080	-0.015	-0.020	0.014	0.058	-0.009	0.007	-0.016
METAPH	0.015	0.100	0.186	-0.179	-0.088	0.042	-0.139	-0.131	0.326
RELIG	0.025	0.097	0.190	-0.184	-0.086	0.046	-0.149	-0.135	0.332
TENTAT	-0.040	0.124	-0.027	-0.001	-0.092	-0.081	0.102	0.050	-0.037
SLEEP	-0.002	-0.012	-0.051	0.021	-0.028	-0.069	0.055	0.027	0.028
DEATH	-0.060	0.045	0.021	-0.015	-0.039	-0.020	0.030	-0.006	0.042
SEXUAL	-0.039	0.074	-0.014	-0.004	-0.053	-0.064	-0.092	0.030	0.054
SCHOOL	0.058	0.028	0.078	-0.060	-0.078	-0.053	-0.011	-0.029	0.041
LEISURE	0.029	0.042	0.066	0.012	-0.016	0.072	-0.036	-0.096	0.089
HOME	-0.005	0.027	0.078	0.006	-0.004	0.107	-0.083	-0.086	0.090
SMILES	0.006	0.050	-0.072	0.007	-0.025	-0.007	0.034	-0.070	-0.016
FEEL	-0.054	0.049	-0.066	-0.026	-0.073	-0.013	0.018	-0.036	-0.030
SPORTS	0.065	-0.021	-0.030	0.073	-0.015	-0.056	0.054	0.005	-0.041

Figure 3: Best LIWC feature selection (accuracy) for each of Schwartz' ten personality value types. The values in the 'Before Feature Ablation' row are based on the full feature set (69 features).

4 Experiments on Automatic Values Identification

We performed several experiments to get a better understanding of the most appropriate linguistic and non-linguistic features for the problem domain. The experiments were designed as a 20 class classification problem, with 'Yes' and 'No' classes for each of the ten Schwartz values. Ten different classifiers were trained, each for a particular value type. Each classifier predicts whether the person concerned is positively or negatively inclined towards the given Schwartz value. The versions implemented in WEKA [Witten and Frank2005] of three different machine learning algorithms were used in the experiments: Sequential Minimal Optimization (SMO; a version of Support Vector Machines, SVM), Simple Logistic Regression (LR), and Random Forests (RF). In all the mentioned experiments the corpora were pre-processed, i.e., tokenized by the CMU tokenizer [Gimpel et al.2011] and stemmed by the Porter Stemmer [Porter1980]. All the lexica were also stemmed in the same way before usage.

4.1 Linguistic Features

LIWC Analysis: LIWC [Pennebaker et al.2001] is a well developed hand-crafted lexicon. It has 69 different categories (emotions, psychology, affection, social processes, etc.) and almost 6,000 distinct words. The 69 categorical features were extracted as user-wise categorical word frequencies. As the text length (for the Essay corpus) or number of messages (Twitter and FB corpora) varies from person to person, Z-score normalization (or standardization) was applied using the equation: $\hat{x} = (x - \mu)/\sigma$, where x is the 'raw frequency count', μ and σ are respectively the mean and standard deviation of a particular feature. After normalizing, each feature vector value is centered around 0 and $\sigma = 1$. This normalization led to an increase in the accuracy figures in many of the cases.

To investigate how each LIWC feature contributes, feature ablation was performed and the Pearson correlations of LIWC features vs value types were analysed as reported in Figure3. The final classifiers were trained using only the features that were contributing for a particular value type. This resulted in a performance boost and also gave reduced time complexity (both model training and testing times). For example, the same accuracy (65.84%) for the Achievement class as obtained by using the full 69 feature set also can be obtained by using only 52 LIWC features. Moreover, the lowest obtained accuracy 53.06% for the Security class increased to 55.80% when considering only 47 features.

4.2 Topic Modeling

In order to find out the bag-of-words features for each value type, i.e., the vocabulary that a person uses more frequently, the MALLET [McCallum2002][3] topic modelling toolkit was used to extract a number of topics. MALLET uses Gibbs Sampling and Latent Dirichlet Allocation (LDA). In a pre-processing stage, stop words were removed and case was preserved. For the Twitter corpus, we tested with different number of topic clusters of sizes 10, 20, 50, 75, and 100, and observed that 50 was the most suitable number. Each of the 50 topics contained an average of 19 words, each with a specific weight attached. The top 5 topics were chosen for each value type, according to these weights, and the words of these topics were added as a new feature set along with the LIWC baseline features.

It was also observed that the rankings of the top 5 topics were almost similar for each Schwartz

[3] http://mallet.cs.umass.edu

Values	Achievement			Benevolence			Conformity			Hedonism			Power		
Classifier	SMO	LR	RF	SMO	LR	RF	SMO	LR	RF	SMO	LR	RF	SMO	LR	RF
LIWC	80.93	80.93	80.10	78.75	78.75	77.38	73.02	72.48	77.93	77.11	76.84	76.02	54.77	50.68	52.59
+Topic	74.66	80.65	80.65	69.21	78.20	77.93	66.76	72.48	73.02	71.66	76.84	76.57	52.32	54.77	51.77
+Lexica	71.10	73.70	69.70	71.90	69.90	65.00	67.20	71.60	68.00	68.00	68.60	60.60	72.80	69.80	59.20
+Non-Linguistic	74.11	80.38	80.93	68.40	78.47	77.38	66.49	72.48	74.11	70.30	76.30	76.57	54.22	55.59	54.22
+Speech-Act	81.10	76.40	68.00	81.00	73.00	66.00	75.00	66.00	66.00	74.00	64.00	63.00	82.00	75.00	63.00

Values	Security			Self-Direction			Stimulation			Tradition			Universalism			Average
Classifier	SMO	LR	RF	SMO	LR	RF	SMO	LR	RF	SMO	LR	RF	SMO	LR	RF	
LIWC	76.29	75.75	74.11	83.38	83.38	75.20	73.57	72.48	70.84	58.04	55.31	55.86	82.02	81.47	80.65	**74.28**
+Topic	70.57	74.93	75.48	76.84	83.38	83.38	64.12	72.47	71.66	52.04	53.95	59.67	74.93	81.47	81.20	**73.70**
+Lexica	70.60	74.30	69.50	75.60	74.40	76.60	68.80	68.60	68.30	73.90	69.50	62.30	78.00	82.20	76.30	**73.38**
+Non-Linguistic	71.18	74.66	75.20	76.57	83.38	83.38	65.58	73.57	71.66	52.59	53.41	55.86	74.39	81.74	82.02	**73.57**
+Speech-Act	78.00	80.00	69.00	78.00	76.00	75.00	73.00	66.00	68.00	80.00	71.00	63.00	89.00	81.10	77.00	**80.00**

Table 1: Automatic Schwartz value detection (accuracy) on the Twitter corpora.

Speech Act	SNO	Wh	YN	SO	AD	YA	T	AP	RA	A	O	Avg.
Distribution	33.37	11.45	15.45	5.16	6.88	15.08	0.41	3.26	0.71	0.07	14.59	
F_1-score	0.45	0.88	0.88	0.72	0.45	0.60	0.72	0.60	0.12	0.77	0.12	0.69

Table 2: Speech act class distributions in the corpus (in %) and speech act classifier performance.

value. The accuracies obtained were almost similar to the accuracies obtained in the previous experiments. However, this time, since the dimension of the feature set is much smaller, the time complexity decreased by almost a factor of 10. Hence the topic modelling was repeated for the social media corpora from Facebook and Twitter, but resulting in a different number of topic clusters, namely 89. Added to the 69 LIWC features this thus resulted in a total of 158 features.

4.3 Psycholinguistic Lexica

In addition to the base feature set from LIWC, two other psycholinguistic lexica were added: the Harvard General Inquirer [Stone et al.1966] and the MRC psycholinguistic database [Wilson1988]. The Harvard General Inquirer lexicon contains 182 categories, including two large valence categories positive and negative; other psycholinguistic categories such as words of pleasure, pain, virtue and vice; words indicating overstatement and understatement, often reflecting presence or lack of emotional expressiveness, etc. 14 features from the MRC Psycholinguistic lexicon were included, namely, number of letters, phonemes and syllables; Kucera-Francis frequency, number of categories, and number of samples; Thorndike-Lorge frequency; Brown verbal frequency; ratings of Familiarity, Concreteness, Imagability and Age of acquisition; and meaningfulness measures using Colorado Norms and Pavio Norms. In order to get these MRC features a machine readable version of it has been used.[4] Feature ranking was done by evaluating the contribution of each feature in an SMO classifier. In addition, the sensorial lexicon Sensicon was used. It contains words with sense association scores for the five basic senses: Sight, Hearing, Taste, Smell, and Touch. For example, when the word 'apple' is uttered, the average human mind will visualize the appearance of an apple, stimulating the eye-sight, feel the smell and taste of the apple, making use of the nose and tongue as senses, respectively. Sensicon provides a numerical mapping which indicates the extent to which each of the five senses is used to perceive a word in the lexicon. We perform feature ablation and also analyze (Pearson) correlations of lexicon features vs values. Finally, classifiers are trained using only the features which are contributing for a particular value.

[4] http://ota.oucs.ox.ac.uk/headers/1054.xml

4.4 Non-Linguistic Features

Social network structure is very useful to predict any person's intrinsic value. For each user in the Twitter corpus, the total number of tweets or messages, total number of likes, average time difference between two tweets/messages, total number of favourites and re-tweets, and their in-degree and out-degree centrality scores on network of friends and followers were used as features adding to a total of 7 features along with the feature set used in the Topic Modelling experiment after observation of the structure of tweets and the previously done linguistic feature experiments. The degree centrality was calculated as of a vertex v, for a given graph *G(V,E)* with $|V|$ vertices and $|E|$ edges, is defined as: $\{C_D = deg(v)\}$.

4.5 Speech Act Features

The way people communicate, whether it is verbally, visually, or via text, is indicative of Per-

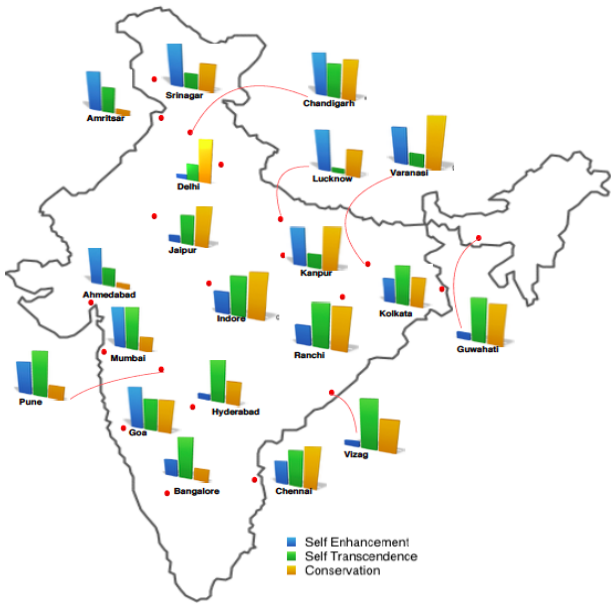
sonality/Values traits. In social media, profile status updates are used by individuals to broadcast their mood and news to their peers. In doing so, individuals utilize various kinds of speech acts that, while primarily communicating their content, also leave traces of their values/ethical dimensions behind. By following the hypothesis of appling2013towards, we have applied speech act features in order to classify personalities/values. However, for this experiment we have restricted speech act classes into 11 major (avoiding 43 fine-grained speech act classes[5]) categories: Statement Non-Opinion (SNO), Wh Question (Wh), Yes-No Question (YN), Statement Opinion (SO), Action Directive (AD), Yes Answers (YA), Thanking (T), Appreciation (AP), Response Acknowledgement (RA), Apology (A) and others (O). A corpus containing 7K utterances was collected from Facebook and Quora pages, and annotated manually. Motivated by the work by [Li et al.2014], we used this corpus to develop an SVM-based speech act classifier using the following features: bag-of-words (top 20% bigrams), presence of "wh" words, presence of question marks, occurrence of "thanks/thanking" words, POS tags distributions, and sentiment lexica such as NRC Linguistic Database (mohammad2013nrc), SentiWordNet [Baccianella et al.2010], and WordNet Affect [Strapparava and Valitutti2004]. The categorical corpus distribution and the performance of the final classifier are reported in Table 2, showing an average F_1-score of 0.69 after 10-fold cross validation.

Automatic speech act classification of social media conversations is a separate research problem altogether, and hence out of scope of the current study. However, although the speech act classifier was not highly accurate in itself, the user specific speech act distributions could be used as features for the psycholinguistic classifiers (resulting in 11 additional features). Experiment on the Twitter Corpus, a noticable performance improvement of 6.12% (F-measure) was obtained.

5 Indian Values Map

Table 3 shows the percentage of the total population of the concerned cities which are inclined to the listed super classes. The percentages were further normalized (with an ε parameter) using the formula-

$$x_{norm} = \frac{x - x_{min} + \varepsilon}{x_{max} - x_{min} + \varepsilon} \times 100 \qquad (1)$$



We observe striking similarities to intuitive perception, i.e., the results obtained resemble our general perception about the city/town. For example, consider the class SC1 referring to Self-Enhancement, the percentage score of Mumbai is the highest which is expected as Mumbai is "*The city of dreams*" for people all over the country who move in to Mumbai to fulfill their ambitions. The huge population in Mumbai leads to immense competition, therefore, it is expected that people of Mumbai are hedonic, power, and achievement oriented. Other progressive cities like Chandigarh, Lucknow and Ahemdabad follow the same trend.

The values of SC2 (openness to change) are equally high for almost all cities, this is a consequence of the fact that a major population in cities are open to changes, they are self directed since they earn their own living and they are also stimulated as they have an active professional and personal life. It is also highly possible, that, since the data has been collected from social-media, the value of SC2 is high, as social-media users belong to the literate section of the society and possess

[5]See for Fine-Gained Speech-Act classes `http://compprag.christopherpotts.net/swda.html`

	AC	BE	CO	HE	PO	SE	SD	ST	TR	UN	SC1	SC2	SC4	SC3	MinT	MaxT	UsersT	AvgT
Ahemdabad	100.0	63.12	100.0	66.18	52.9	70.4	100.0	100.0	56.16	100.0	59.74	100.0	52.39	47.41	2.0	3249.0	136.0	1922
Amritsar	58.33	61.86	100.0	29.27	64.38	20.0	100.0	100.0	5.0	100.0	71.94	100.0	10.0	46.0	2.0	3250.0	201.0	2109
Bengalaru	19.16	81.5	100.0	40.72	50.74	42.0	100.0	100.0	16.36	100.0	31.85	100.0	21.59	73.55	2.0	3250.0	206.0	1846
Chandigarh	100.0	60.57	100.0	71.45	69.13	64.4	100.0	100.0	52.99	100.0	90.17	100.0	1.01	43.78	2.0	3250.0	113.0	1864
Chennai	58.33	63.36	100.0	56.36	49.01	100.0	100.0	100.0	50.23	100.0	33.28	100.0	53.34	47.75	1.0	3248.0	200.0	1698
Delhi	60.83	44.8	100.0	0.18	52.28	43.6	100.0	100.0	47.6	100.0	10.0	100.0	50.88	21.35	2.0	3249.0	212.0	1991
Goa	100.0	85.49	100.0	100.0	40.3	12.8	5.0	100.0	11.64	100.0	56.73	100.0	43.15	43.15	4.0	3248.0	91.0	1942
Guwahati	31.66	86.63	100.0	45.27	36.54	67.2	100.0	100.0	43.23	100.0	12.36	100.0	64.14	71.04	1.0	3250.0	121.0	1656
Hyderabad	60.0	76.99	100.0	22.72	44.44	60.8	100.0	100.0	17.5	100.0	5.0	100.0	19.63	64.58	1.0	3248.0	205.0	1880
Indore	100.0	62.64	100.0	40.72	49.19	100.0	31.19	100.0	48.28	100.0	28.05	100.0	51.51	46.73	1.0	3249.0	138.0	1634
Jaipur	100.0	59.15	100.0	11.09	54.32	100.0	100.0	100.0	50.23	100.0	9.66	100.0	53.34	41.77	22.0	3247.0	68.0	2161
Kanpur	100.0	39.03	100.0	5.0	100.0	100.0	100.0	100.0	34.81	100.0	75.27	100.0	46.52	25.97	2.0	3247.0	168.0	1920
Kolkata	100.0	86.42	100.0	54.18	54.93	100.0	100.0	100.0	57.71	100.0	54.35	100.0	60.35	80.22	1.0	3250.0	470.0	1702
Lucknow	100.0	5.0	100.0	54.54	80.37	56.8	100.0	100.0	61.75	100.0	86.37	100.0	52.27	10.0	3.0	3249.0	280.0	1940
Mumbai	**100.0**	100.0	100.0	**58.0**	68.39	41.6	55.04	100.0	33.87	100.0	**100.0**	100.0	34.91	**100.0**	8.0	3248.0	205.0	2032
Pune	100.0	94.59	100.0	58.0	53.7	5.0	100.0	100.0	29.36	100.0	70.68	100.0	27.71	92.56	2.0	3248.0	206.0	1915
Ranchi	5.0	58.61	100.0	33.09	48.58	52.0	100.0	100.0	34.81	100.0	19.8	100.0	38.88	41.0	4.0	3250.0	83.0	1710
Srinagar	37.5	44.47	100.0	25.45	70.8	70.0	100.0	100.0	58.24	100.0	58.32	100.0	60.85	20.88	2.0	3246.0	136.0	1715
Varanasi	100.0	48.91	100.0	71.63	54.25	100.0	100.0	100.0	100.0	100.0	78.28	**100.0**	100.0	27.21	2.0	3250.0	152.0	1557
Vizag	100.0	93.81	100.0	40.18	5.0	100.0	100.0	100.0	37.37	100.0	8.55	100.0	53.52	81.5	4.0	3247.0	68.0	1964

Table 3: City-Wise Indian Values

	AC	BE	CO	HE	PO	SE	SD	ST	TR	UN	SC1	SC2	SC4	SC3	MinT	MaxT	UsersT	AvgT
North India	37.83	15.0	100.0	15.0	100.0	45.53	100.0	100.0	100.0	100.0	100.0	100.0	100.0	33.0	2.0	3250.0	1466.0	1911
South India	15.0	91.36	100.0	22.58	15.0	93.75	100.0	100.0	15.0	100.0	33.0	100.0	10.0	89.96	1.0	3250.0	679.0	1824
East India	18.91	98.7	100.0	55.48	45.92	100.0	100.0	100.0	70.15	100.0	63.5	100.0	66.0	97.19	1.0	3250.0	674.0	1694
West/Central India	**100.0**	100.0	100.0	100.0	57.97	15.0	15.0	100.0	29.09	100.0	87.22	33.0	12.55	100.0	1.0	3249.0	640.0	1896

Table 4: Region-Wise Indian Values

a good educational background. As there is no differentiation of SC2 values we are not reporting the values of SC2 in Indian Values Map. In SC3, which refers to Self-Transcendence, Mumbai again scores the highest, with inhabitants from all over the country, Mumbaikars, are prone to be universal, since they accept people from all backgrounds. They also tend to be more benevolent and warm hearted, this trend can be observed in all other rapidly developing cities like Bangalore, Kolkata, Vizag, Indore etc. In the case of SC4, which refers to conservation, the temple town, Varanasi tops among others. This result can be easily anticipated since the Varanasi is high on tradition, individuals conform to rules in fear of god and yearn for a secure living. It can also be observed that rapidly-developing cities like Chandigarh, Bangalore, Hyderabad, Pune, Amritsar have low values, indicating the diminishing tradition and conformity to ancient rules and regulations in these cities. The trends have been visually demonstrated in Figure 4, indicating the location of each city and the respective percentage of population possessing the class of values as defined by the 4 super-groups.

For analysis of results at regional level, India can be broadly classified into four regions namely, North, South, East and West India. The region-wise results are aggregated and reported in the Table 4.

The results obtained at regional level are also close to general intuition about these regions which will be clear from the derived implications discussed next. One of the major implications is that West India tops in terms of achievement (AC) oriented and hedonistic (HE) people. In general, North Indians are among the most power (PO) oriented people. For example, majority of politicians and administrative officers are from North India as compared to other parts of India. In terms of traditional people, North India comes at top. This result can be attributed to the fact that the analysis of North India is based on more traditional cities like Srinagar, Kanpur, Lucknow and Varanasi. These trends are clearly reflected in Table 3 & 4.

6 Discussion and Conclusion

In this paper, we propose a computational Schwartz values model and applying to various Indian users and finally created a Indian values map.

As can be seen from the result that a few Schwartz values such as Self-Direction and Security are relatively difficult to identify, while on the other hand the accuracies for certain value types such as Power and Tradition are persistent and seem to be more salient. The results also indicate that social media text is difficult for automatic classification, which is obvious from its terse nature. A major limitation is that the collected social network corpus is skewed, therefore the results do not have much deviations from the majority baselines. Since the classes are imbalanced, the system is always statistically biased towards the major class. The expected solution is having more

data, and we are collecting more data, aiming to reach data from 10,000 users. The collected data will be publicly released to the research community.

For the machine learners, closer analysis revealed that SMOs performance was irregular and random, which might be an indication of over-fitting. On the other hand, the performance of the Random Forests classifier increased when the number of features was increased, resulting in a larger forest and hence for most value types Random Forests performed better than the other two classifiers with less over-fitting.

We are also very keen on the applied side of this kind of models. Presently we are analysing the community detection problem in social media in relation to values. Another interesting application could be comparative societal analysis between the Eastern and Western regions of the world. Relations among personality and ethics could also be explored.

References

[Agle and Caldwell1999] Bradley R. Agle and Craig B. Caldwell. 1999. Understanding research on values in business a level of analysis framework. *Business & Society*, 38(3):326–387.

[Argandoña2003] Antonio Argandoña. 2003. Fostering values in organizations. *Journal of Business Ethics*, 45(1–2):15–28.

[Baccianella et al.2010] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May. European Language Resources Association (ELRA).

[Bienaymé1853] Irénée Jules Bienaymé. 1853. Considérations à l'appui de la découverte de Laplace sur la loi de probabilité dans la méthode des moindres carrés. *Comptes Rendus de l'Académie des Sciences Paris*, 37:309–324.

[Clawson and Vinson1978] C. Joseph Clawson and Donald E. Vinson. 1978. Human values: A historical and interdisciplinary analysis. *Advances in Consumer Research*, 5(1).

[Gimpel et al.2011] Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 2011. Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 42–47. Association for Computational Linguistics.

[Henrique2015] Jefferson Henrique. 2015. https://github.com/jefferson-henrique/getoldtweets-java.

[Hofstede et al.2010] Geert Hofstede, Gert Jan Hofstede, and Michael Minkov. 2010. *Cultures and Organizations: Software of the Mind*. McGraw-Hill Education, 3 edition.

[Hood2003] Jacqueline N Hood. 2003. The relationship of leadership style and CEO values to ethical practices in organizations. *Journal of Business Ethics*, 43(4):263–273.

[John and Srivastava1999] Oliver P. John and Sanjay Srivastava. 1999. The big five trait taxonomy: History, measurement, and theoretical perspectives. volume 2, pages 102–138. Guilford, New York, NY.

[Kahle et al.1986] Lynn R. Kahle, Sharon E. Beatty, and Pamela Homer. 1986. Alternative measurement approaches to consumer values: The list of values (lov) and values and life style (vals). *Journal of Consumer Research*, pages 405–409.

[Li et al.2014] Jiwei Li, Alan Ritter, Claire Cardie, and Eduard Hovy. 2014. Major life event extraction from twitter based on congratulations/condolences speech acts. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1997–2007, Doha, Qatar, October. Association for Computational Linguistics.

[McCallum2002] Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit.

[Pennebaker et al.2001] James W. Pennebaker, Martha E. Francis, and Roger J. Booth. 2001. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*, 71:2001.

[Porter1980] M.F. Porter. 1980. An algorithm for suffix stripping. *Program*, 14(3):130–137.

[Rokeach1973] Milton Rokeach. 1973. *The nature of human values*. Free Press, New York, NY.

[Schwartz et al.2001] Shalom H. Schwartz, Gila Melech, Arielle Lehmann, Steven Burgess, Mari Harris, and Vicki Owens. 2001. Extending the cross-cultural validity of the theory of basic human values with a different method of measurement. *Journal of Cross-Cultural Psychology*, 32(5):519–542.

[Schwartz2012] Shalom H Schwartz. 2012. An overview of the Schwartz theory of basic values. *Online Readings in Psychology and Culture*, 2(1):11.

[Stone et al.1966] Philip J. Stone, Dexter C. Dunphy, and Marshall S. Smith. 1966. *The General Inquirer: A Computer Approach to Content Analysis*. MIT press.

[Strapparava and Valitutti2004] Carlo Strapparava and Alessandro Valitutti. 2004. WordNet Affect: an affective extension of WordNet. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC-2004)*, Lisbon, Portugal. European Language Resources Association (ELRA).

[Tchébichef1867] Pafnuty Lvovich Tchébichef. 1867. Des valeurs moyennes (translated into French by N.V. Khanykov). *Journal de Mathématiques Pures et Appliquées*, 12(2):177–184.

[Wilson1988] Michael Wilson. 1988. MRC psycholinguistic database: Machine-usable dictionary, version 2.00. *Behavior Research Methods, Instruments, & Computers*, 20(1):6–10.

[Witten and Frank2005] Ian H. Witten and Eibe Frank. 2005. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

[Yamamoto2014] Yusuke Yamamoto. 2014. Twitter4j-a java library for the twitter api.