An Empirical Model of Gender Based Human Trait Identification from Blogs

Amit Sinha
Department of Information Technology
ABES Engineering College
Ghaziabad, India
amitsinha@abes.ac.in.

Dr. Ashok K. Sinha, Member IEEE
Department of Computer Science & Engineering
Greater Noida Institute of Technology
Greater Noida, India
aksinha@computer.org

***Abstract*— The growth of blogosphere has prompted researchers for knowledge discovery from the blogs. This paper presents a brief review on research works in this area and proposes an empirical model of gender and trait identification. The feature vector comprises of three elements viz., formal measure, gender preferential measure and stylistic measure. The various significant parts-of-speech found in the blogs are used in computing the features. The model is implemented in two stages viz. author gender identification and their personality trait identification. The model identifies gender based traits viz., Extrovert, Introvert and Thinker. The model is trained satisfactorily with textual data collected from blogs using adaptive network fuzzy inference system algorithm.**

Keywords-component; Gender; Personality; Machine Learning; Use error; Adaptive systems

I. INTRODUCTION

The research interest in the area of cognitive science has been growing over the last few decades for identification of author's personality trait from their blogs in social media.

Human activities, such as social interactions, writing blogs and gathering information mediated by digital services and devices are growing. Although people may choose not to reveal certain personal information, much of information on author's trait can be extracted from data available from digitally mediated activities [1-4].

Human personality reflects certain traits which are common among different people. Various researchers working in the domain of human trait theory have carried out studies on predicting human behavioral characteristics by establishing relationship with trait and its attribute. Gordon Allport [5], an earliest researcher worked on over 4,000 listed personality traits. His work gave few categorization of traits such as dominating traits (cardinal traits) which develop from childhood or by birth, individual traits (central traits) which include personal behavior such as honest, anxious and happy and it varies from person to person and secondary traits that develop with experience and it varies with persons. Later Raymond Catell [6] analyzed this list using factor analysis and reduced it to 16 key personality factors. Each person contains all of these 16 traits to a certain degree but they may be high in some traits and low in others. A different characterization was done by Hans Eysenck [7, 8] such as introversion, extroversion, neuroticism, psychoticism. The Big five Theory of Personality was coined by Lew Goldberg and later its factor analysis was made by McCrae and Costa. This theory considers that there are five core personality traits common to humans. These are Openness, Conscientiousness Extraversion (Extroversion and Introversion), Agreeableness, and Neuroticism, called OCEAN. The OCEAN traits are the basic and fundamental traits of a person.

The application of lexical approach to trait analysis led to find big five traits in languages [9]. The approach is to determine the fundamental personality traits by analyzing language. A trait adjective that possesses many synonyms is more likely to represent a more fundamental trait than that with few synonyms used in the text in any language. Bloggers may have different emotions and use varying length of adjectives. They also follow various short words, understandable blog words and emoticons etc.

Human traits are found to be dependent on various part-of-speech (POS) contained in the author's text. Different set of POS are used by men and women. The POS extracted from text can be used as input to computational model of personality traits of men and women. It has been found that personality traits of men and women differ partly because of temperamental or hormonal difference, yet another reason could be their gender role in the society. Two meta-analyses of Costa et al. [2] studied on personality traits among men and women. It has been observed that the women generally use soft words [10, 11]. They score higher on anxiety, trust and tender-mindedness. Men on the other hand use more assertive words in their writings. Although the recorded textual data contain much direct information about the author, but greater challenge lies in its lexical and semantic analysis for estimating the gender and the personality trait of the author. In recent times we have seen an emergence of social media and its uses. The blogs in the media contain much of information which can be used for behavioral studies .Several empirical studies have been carried out by researchers, yet a structured computational model need to be developed in a formalized manner. With this objective the present paper has proposed an empirical model for gender and personality trait classification using features extracted from textual data in the blogs. The model is developed in two stages, firstly gender identification and secondly trait identification. It is based on machine learning approach using adaptive network fuzzy inference system [3].

978-1-4799-6023-1/15 $31.00 © 2015 IEEE
DOI 10.1109/CICT.2015.125

II. A BRIEF REVIEW ON RELATED EMPIRICAL WORK

In the present age usage of social media is increasing with upsurge in the availability of textual data. Such data contain much of information about the author's behavior .This has prompted researchers working in the area of behavioral sciences and computational modeling for knowledge discovery and predict the individual behavior in the social media environment. In order to provide an empirical solution to this problem Arjun Mukherjee et al [12] proposed a technique using a feature based selection method for gender classification of blog authors. The proposed feature set contains few parts-of-speech (POS) which are tagged on the text. In further extension to this work few more significant POS tags like conjunction and determiner can be included in the feature set to improve the classification. Schler et al [13] analyzed a corpus of tens of thousands of blogs and indicates significant differences in writing style and content between men and women bloggers as well as among authors of different ages.

Such differences can be exploited to determine an unknown author's age and gender on the basis of a blog's vocabulary. The style of writing may be a significant feature in gender classification. Cheng et al. [14] examined the text submitted by the author and identified the gender based on their style of writing. The paper concluded that men and women use different style of writing. A feature set is built for this identification problem [15, 16]. Machine learning algorithms are used on the proposed features. In further research in text mining and sentiment analysis, Hamouda and Akaichi [17] illustrated Tunisian users' statuses on "Facebook" posts during the "Arabic Spring" era. They found that attributes of sentiments differ from men to women; therefore, there should be different technique for individual gender. Another empirical work by Ansari Y Z, Azad A B, Akhtar Halima [18] on text mining of blogs used frequency counter, term frequency and inverse document frequency for characterization of blogs. Naive Bayes probabilistic model has been used for machine learning.

III. THE PPROPOSED MODEL

A review on the past works has prompted the authors of this paper to develop an intelligent input-output model of trait identification with gender difference based on the texts in the blogs. The model extracts a feature vector from textual data available in social media which acts as input to the model. The output of the model is obtained by logical aggregation of inference rules by biological-inspired neural network. It first classifies the gender and then their personality traits. It assumes a fuzzy inference rules for classification which is modified by using Adaptive Network Fuzzy Inference System in each training step. The data available from social media are used in two phases of machine learning viz. training and validation.

IV. HYPOTHESIS

The model formulation in the present work is based on the following hypotheses:

(i) Human gender and trait may be identified by the feature extracted from the text written by author.

(ii) The feature vector comprises of three elements, viz., formal measure, gender preferential measure and stylistic measure which are computed from part-of-speech extracted from text.

(iii) The feature vector acts as input to ANFIS model which identifies the gender and the trait as outputs. The ANFIS generates fuzzy inference rules for each model.

V. MODEL DEVELOPMENT

The model formulation is based on the above hypothesis. Its architecture is that of ANFIS as fusion of neural network and fuzzy inference system linked together in an adaptive manner.

A. Model Formulation

The proposed model comprises of two functions FG (Gender) and FT (Trait). The function FG is used to classify the gender, Fig. 1, while FT is used to identify the trait of the author, Fig. 2.

Gender Model: The model equation is given as:

$$FG = f(FM, GPM, SM) \quad (1)$$

FIGURE 1. GENDER MODEL

The various model variables are defined as:

A.1 Formal Measure (FM)

The probability of using noun, adjectives, preposition and articles used by men in their text is higher than other parts of speech while the probability of using pronoun, verb, adverb and interjection by women is higher than other parts of speech. The proposed function of FM [12] includes the following:

(i) Parts-of-speech noun, adjective, preposition, conjunction, pronouns, verbs, adverbs and interjections.

(ii) On analyzing several blogs, it is found that blogs have more 'determiners' like 'the, that, my, a, an etc' than 'article'. Therefore, in formulation of FM, determiner is included.

(iii) The final expression, is multiplied by 0.4. This value could be in the range of 0.4 to 0.5. Machine learning is found to be better at this value.

Hence in this paper, FM is defined as:

$$FM = 0.4 * [(NN + JJ + IN + DT + CC) - (PN + VB + RB + UH) + 100] \quad (2)$$

NN is total Noun, JJ is total Adjectives, IN is total Preposition, DT is total Determinant, CC is total Conjunction, PN is the total number of pronouns, VB is total number of Verbs, RB is total number of adverbs and UH is the total number of Interjection present in posted blogs.

A Part-Of-Speech Tagger (POS Tagger) is used for the extraction of the text. The POS tagger is software that reads text and assigns parts of speech to each word [19, 20]. A distinct tag is assigned to each POS word.

A.2 Gender Preferential Measure (GPM)

GPM includes several features like negativity, positivity, anger, certainty emotions found in the blog. These attributes are assigned with some threshold values to categorize the gender men and women. The threshold values are determined on the basis of analysis of a large database of blogs of known authors.

A database is created to classify different parts-of-speech into different classes of emotions.

The average value of all threshold values, denoted by %GPM, is being taken for the gender classification and it is found different values for men and women.

A.3 Stylistic Measure (SM)

This feature captures author's writing styles. The style of writing a blog may contain the following: abbreviations, blog words and emoticons like "hmm", "thnx" and smiley. The women generally use more stylistic words than men.

Trait Model: The model equation is given as**:**

$$FT = f(FPP, SPP, TPP, JJ, NN, DT, VB) \quad (3)$$

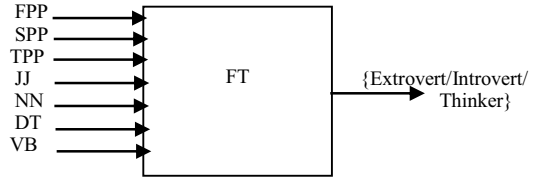

FIGURE 2. TRAIT MODEL

The variables NN, DT and VB are defined in gender model. The count of total pronouns PN is classified into FPP (first person pronoun), SPP(second person pronoun) and TPP(Third person pronoun). The count of total adjective JJ is divided into positive and negative adjectives.

B. Model Architecture

The model architecture is that of an Adaptive Network Fuzzy Based Inference System [3] comprising of five layers with nodes in each of the layers viz., input layer, input membership function layer, inference rule layer, output membership function layer and output layer. In the process of machine learning it adaptively modifies fuzzy relationship between input and output at each epoch. In the present work three input nodes and one output node for gender model have been configured for machine learning. The ANFIS architecture used in the gender model is shown in Fig. 3. This network architecture is similar in the trait model with the number of input nodes being modified to seven.

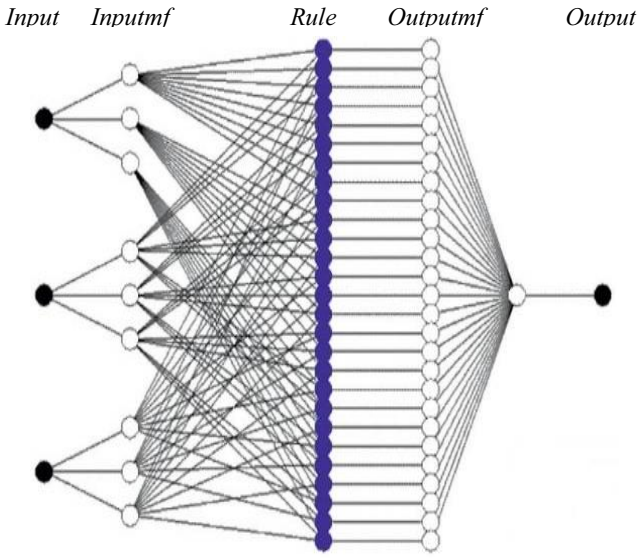

FIGURE 3. ANFIS ARCHITECTURE WITH THREE INPUTS AND ONE OUTPUT

VI. METHODOLOGY

The work has been carried out in the following steps:

1. Collection of texts from blogs in social media.
2. Gender classification
   - 2.1 Calculation of FM using POS tagger (section V. A.1).
   - 2.2 Calculation of GPM (section V.A.2).
   - 2.3 Calculation of SM (section V.A.3).
   - 2.4 Training of the gender model using ANFIS.
   - 2.5 Testing the gender model with additional data set.
3. Trait Identification
   - 3.1 Extraction of the attributes as per the Table III using POS tagger.
   - 3.2 Training the trait model using ANFIS.
   - 3.3 Testing the trait model with additional data set.

VII. IMPLEMENTATION AND RESULT

Step-1: Creating database of texts from blogs

A set of one hundred blogs for training and twenty blogs for testing have been collected from websites www.blogspot.com, www.indianblogs.com, www.indianwomenblogs.com.

The calculation of various features from the blogs is given below. A sample text is taken from www.mkgandhi-

sarvodaya.org / articles / mgjournalism.htm. and www.hindustantimes.com/StoryPage/ Print/15885.aspx. The blog is written by the great Mahatma Gandhi in Young India on 25 July 1925.

"I have taken up journalism not for its sake but merely as an aid to what I have conceived to be my mission in life. My mission is to teach by example and precept under severe restraint the use of the matchless weapon of satyagraha which is direct corollary of non-violence and truth. I am anxious, indeed I am impatient, and to demonstrate that there is no remedy for the many ills of life save that of non-violence. It is a solvent strong enough to melt the stoniest heart. To be true to my faith, therefore, I may not write in anger or malice. I may not write idly. I may not write merely to excite passion. The reader can have no idea of the restraint I have to exercise from week to week in the choice of topics and my vocabulary. It is training for me. It enables me to peep into myself and make discoveries of my weaknesses. Often my vanity dictates a smart expression or my anger a harsh adjective. It is a terrible ordeal but a fine exercise to remove these weeds."

A sample of calculation for the above text is shown below.

Step-2: Gender Classification

Step-2.1: Calculation of FM using POS tagger

The FM for above blog is calculated on the basis of total number of defined attributes present in blog.

NN =41, JJ =15, IN =20, DT = 21, CC=8, PN = 22 (FPP=17, SPP =0, TPP=5), VB=29, RB=11, UH=0 . (4)

The FM is calculated using (2)

F = 0.4 * [(41 + 15 + 20 + 21 + 8) – (22 + 29 + 11 + 0) + 100] = 57.2 .

The calculation of FM for several blogs shows that for men it is generally higher than value 50 while for women it is less than 50.

Step-2.2: Calculation of GPM

The sample text contains few emotion words for GPM as per the section IV A.2. The words present in the sample text such as '*impatient', 'non-violence'* and *'ills'* come under negative emotions*, 'true'* comes under positive emotions*, 'anxious'* and *'anger'* come under anger emotions.

The total number of GPM words in the sample text is 6 out of total words of sample text 185. Thus, the %GPM is 3.24%. On analyzing several blogs the value of %GPM lies in between 0 to 15% for men while it is greater than 15% for women.

Step-2.3: Calculation of SM

Analyzing the sample text, no stylistic words are found i.e. the number of SM= 0. The analysis of several blogs show that for the men, value of SM is less than or equal to 7% while it is greater than 7% for women.

Step-2.4: Training of the gender model using ANFIS

The calculated feature set from few samples for training data are shown in Table I and Table II.

TABLE I. TRAINING DATA FOR MEN

FM	GPM	SM
40	16	9
45	20	10
30	18	11
25.5	19.2	12
10	25	10
5	18	11
22	18	14
25	21	14
36.2	16.6	9
15	19	16
21	18	17
15	19	10
16	16	10
25	17	17
10.4	18.3	10
17.2	16.4	10
37	16	10
45	17	12
48	19	12
30.2	18.4	12

TABLE II. TRAINING DATA FOR WOMEN

FM	GPM	SM
60	10	5
55	15	6
56	18	7
54	11	3
80	5	7
57.2	3.24	0
52	8.5	4
51	6	4
60	12	6
53	10.6	6
55.4	12	7
50	11	7
60.2	10.5	6
50.4	11	7
55	6	7
51	12	7
53	9.5	7
55	10.4	6
51.2	12	6
50	12	4

The fuzzy inference rules are adaptively modified according to errors in the training process. The implementation of ANFIS is done in MATLAB and the resulting error for men is 1.3111e-05 and for women it is 1.7639e-04.

Step-2.5: Testing the gender model with additional data set

The testing of the gender model was performed using additional set of data and the result was found to be correct for 86.56 % of the test cases.

Step-3: Trait Identification

For each gender, this paper identifies the trait of author. Three traits, viz., ***Extrovert, Introvert and Thinker*** are taken for the purpose. The same set of data on blogs is used. The various features as inputs to trait model in (3) are calculated. A sample of calculation for the blog as stated in section VI step 1

is given below; here PN = FPP + SPP +TPP. These steps are used to identify the trait of the individual author.

Step-3.1: Extraction of the attributes as per the Table III using POS tagger

The different types of attributes in the sample text are calculated in percentage as shown below:

In the sample text, total words =185.
The computation of attributes as per the Table III is

%FPP =9.19, %SPP =0, %TPP=2.7, %JJ= 8.1, %NN =22.1, %DT=11.35, %VB= 15.67. (5)

Step-3.2: Training the trait model using ANFIS

A sample of ten data sets for male with 'Thinker' trait is shown in Table III.

TABLE III. INPUTS OF TRAIT MODEL FOR 'THINKER' MEN

FPP	SPP	TPP	JJ	NN	DT	VB
7.12	1.5	3	2.3	16	6.2	12.3
9.19	0	2.7	8.1	22.1	11.35	15.67
8.16	2	1.4	3.6	16	8	11.6
9	1	2	3	17	9	12
10	1	1	1	17	6	14
8.5	2	1.5	1	15.67	6.5	11.67
9	1	2	1	16	7	10.5
10	1	3	3.5	15	7.5	11.67
11	1	1	1	16	8	12
7.5	2.6	2	2.5	15.67	7	12.67

Similar tables are prepared for different traits like extrovert and introvert. The whole data set thus prepared are the inputs for ANFIS model for training. The resulting error for 'Thinker category' of men is 1.0755e-05 and the model errors for other traits are in the same range.

Step-3.3: Testing the trait model with additional data set

The testing of the trait model was performed using additional set of data and the result was found to be correct for 85.15 % of the cases.

VIII. CONNCLUSION

The present work has successfully developed an intelligent model for author's gender identification and his/her personality trait classification from features extracted from blogs. While writing blogs it is assumed that the author is free from any environmental influences. The model is based on the assumption that men and women differ significantly in their use of part-of-speech, emotions and stylistic feature like use of emoticons in the blogs. Besides, human personality trait has been found to have influence on the use of adjectives and pronouns used in his/her writing. This has been identified from the part-of-speech of the text. The input–output model is trained with data collected from blogs using ANFIS on MATLAB. The converged value of error is as low as 1.7639e-04 for gender classification and 1.0755e-05 for trait classification. While testing we could achieve 86.56% accuracy for gender classification and 85.15% accuracy for trait classification. By designing appropriate psycho-linguistic and gender-linked features, we observe that word-based features and structural features play important roles in gender and personality trait classification.

REFERENCES

[1] Goldberg, L.R.: The structure of phenotypic personality traits. American Psychologist 84, 26–34 (1993).

[2] Costa, P.T. Jr,Terracciano A, McCrae R (2001): Gender Differences in personality traits across culture, Journal of Personality & Social Psychology, 81,322-331.

[3] Jyh-Shing, Roger Jang (1993): ANFIS: Adaptive Network Based Fuzz Inference System, IEEE transaction on SMC vol 23, No. 3, May/June 1993, pg. 665-685.

[4] Argamon, S., Koppel, M., Pennebaker, J. W., Schler, J. 2007. Mining the Blogosphere: Age, Gender and the varieties of self-expression, 2007.

[5] Allport F.H. & Allport G.W(1921). Personality Traits: Their Classification and Measurement.

[6] Cattell, R. B. (1990). Advances in Cattellian personality theory. In L. A. Pervin (Ed.), Handbook of personality: Theory and research (pp. 101-110). New York: Guildford.

[7] Eysenck, H.J.: The Measurement of Personality. Medical and Technical Publishers, Lancaster (1976).

[8] Eysenck, H. J., & Eysenck, M. W. (1985). Personality and individual differences: A natural science approach. New York: Plenum.

[9] Gordon W Allport and Henry S. Odbert., 1936 "Trait-Names :A Psycho-Lexical Study", Published by Psychological Review Company.

[10] Fiengold, A (1994) Gender Differences in Personality-a meta-analysis, Psychological Bulletin.

[11] Herring, S. C., & Paolillo, J. C. 2006. Gender and genre variation in weblogs, Journal of Sociolinguistics,10 (4), 439-459.

[12] A. Mukherjee and B. Liu, "Improving Gender Classification of Blog Authors" in the Proceedings of 2010 Conference on Empirical Methods in Natural Language Processing pg. 207-217, ACM.

[13] Jonathan Schler, Moshe Koppel, Shlomo Argamon and James Pennebaker, "Effects of Age and Gender on Blogging", ACM 2010.

[14] Na Cheng, R. Chandramouli and K.P. Subbalakshmi "Author gender identification from text" ACM 2011.

[15] Agrawal, R. and Srikant, R. 1994. Fast Algorithms for Mining Association Rules. VLDB. pp. 487-499.

[16] Blum, A. and Langley, P. 1997. Selection of relevant features and examples in machine learning. Artificial Intelligence, 97(1-2):245-271.

[17] Hamouda and Akaichi,"Social Networks' Text Mining for Sentiment Classification: The case of Facebook' statuses updates in the Arabic Spring Era", International Journal of Application or Innovation in Engineering & Management (IJAIEM), ISSN 2319 – 4847, Vol. 2, Issue 5, May 2013 pp. 470-479.

[18] Ansari Y Z, Azad A B ,Akhtar Halima "Gender Classification of Blog Authors, "International Journal of Sustainable Development and Green Economics (IJSDGE), ISSN No.: 2315-4721, V-2, I-1, 2, 2013 .

[19] Kristina Toutanova and Christopher D. Manning (2000): Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger, In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000), pp. 63-70.

[20] Kristina Toutanova, Dan Klein, Christopher Manning and Yoram Singer (2003) Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network, In Proceedings of HLT-NAACL 2003, pp. 252-259.