An Improved Human Trait Modeling using Fuzzy Inference System

Amit Sinha
ABES Engineering College
Ghaziabad, India

Ashok K. Sinha
ABES Engineering College
Ghaziabad, India

Abstract

This paper presents a significant improved system in the field of trait identification. The model designed in this work analyzes the blogs or input text and classifies the traits into five major categories viz. Neuroticism, Extraversion, Openness, Conscientiousness and Agreeableness. The blog or text is first tagged through POS tagger then a feature vector matrix [FVM] is generated according to the attributes of the trait chart. Each column of FVM is calculated in its domain that improves the final result of trait identification. The FVM is then implemented through Fuzzy Inference System [FIS] on MATLAB 7.0 software. The result of the proposed model is improvement over similar work by other researchers [1, 2, 3]. This model has various applications like predicting behavior, creating team for the same project and counseling the students.

Keywords

Trait, POS Tagger, Feature Vector Matrix, Fuzzy Inference System, MATLAB tools

1. Introduction

Personal blog or online diary is one of the famous category in which the authors write their thoughts, feelings and communicate with other people. Some blogs focus on a particular topic such as news blogs, political blogs and movie blogs etc. In recent years, several researchers have been working on the classification of blog authors using different features such as content words, dictionary based content analysis, parts of speech tags and feature selection along with a supervised learning algorithm [1-5].

In Gordon Allport's approach, central traits are basic to an individual's personality whereas secondary traits are more peripheral. Common traits are those recognized within a culture and may vary between cultures. Cardinal traits are those by which an individual may be strongly recognized. Since Allport's time, trait theorists have focused more on group statistics than on single individuals. Allport called these two as "nomothetic" and "idiographic" respectively. There are a nearly unlimited number of potential traits that could be used to describe personality. The Statistical technique of factor analysis, however, has demonstrated that particular clusters of traits reliably correlate together. Eysenck has suggested that personality is reducible to three major Traits. Other researchers argue that more factors are needed to adequately describe human personality. Many Psychologists currently believe that five factors model is sufficient. This model classifies the traits as Neuroticism, Extraversion, Openness, Conscientiousness and Agreeableness. These traits are analyzed on several attributes. Virtually all trait models Extraversion Vs Introversion as a central dimension of human trait. Another prominent trait that is found in nearly all models is Neuroticism or emotional instability.

The author identification involves various calculations and can not be concluded directly but it may capture some uncertainty in human perception. On the basis of uncertainty, the traits can have lower or higher values.

2. Review on Related Research Work

Carlo Strapparava and Rada Mihalcea [1] used the text with focusing on the emotion classification of news headlines extracted from news websites. They analyzed for the automatic annotation of emotion in text. They conducted an inter tagger agreement for the emotions viz. anger, disgust fear, joy, sadness and surprise. The text was analyzed on the basis of words referring to direct emotional states (e.g. happily) called direct affective words and referring indirect emotion states (e.g. threatening or killer) called indirect affective words. The analysis was done only for emotion classification and concentrated on emotional word. The feature set contains only emotional words and can be used to say anger authors or surprise news etc. However, in this work, the identification of trait is not adequate. The work can be enhanced by addition of more traits. I*n the work presented in our paper, many more emotional words such as positive adjective words (PAW) and negative adjective words (NAW) are also grouped together and then tagged all.*

The gender classification was studied by Arjun Mukherjee and Bing Liu [2]. They proposed two novel techniques:- POS sequence patterns and EFS algorithm [pg 212] to improve the previous performance. The gender classification is experimented on the basis of such studies that women's language makes more frequent use of emotionally intensive adverbs and adjectives like beautifully, smartly and is more punctuated while men's language is more proactive at solving problems. The work showed women use generally more PAW while men use average combination of PAW and NAW in their blogs. *In our paper, we have also considered the Noun words (NW) with PAW and NAW. This yields improved performance of the model.*

Haytham Mohtassab and Amr Ahmed [3] worked on online diaries by mining their diaries text. The contents written b y bloggers are analyzed to identify its nature and properties. The work is performed on several texts of the same author and then selected the best set of features that improved the identification percentage. The experimental works starts with text collection then feature extraction then creation of feature vector Matrix (FVM). The work is performed on six different authors, five different post count per author and eleven different post lengths. This makes 330 groups in total. Each experiment group is repeated 150 times. Then the total FV is passed to support vector machine (SVM) which is used as classification algorithm. The large corpus requires more analysis and time. Also the outcome is declared on the basis of all the texts collected and the work gives 90% accurate result. *Our work takes individual text independently and gives the result in percentage of perception using fuzzy inference system.*

Jon Oberlander and Scott Nowson [4] investigated the accuracy when classifying authors on four important personality traits: Neuroticism, Extraversion, Agreeableness and Conscientious. Weblogs are explored both binary and multiple classification using different set of n-grams features. This work is done to achieve more finer grained multiple classifications. The bloggers may be more or less extravert and not as most extravert because they express themselves in public and they write in their diaries before posting it. Same pattern is applied for three other traits. *We, in our work, considered one extra measurement of trait i.e. average trait along with less and more personality traits. This helps in finding the average trait of authors of the blogs.*

Scott Nowson and Jon Oberlander [5] showed some improved performance in authorship identification. This time they considered 5-point Likert scale i.e. Neuroticism, Extraversion, Openness, Agreeableness and Conscientious. Here they worked with two separate corpora of weblogs- original corpus (OC) and new corpus (NC). The language model, introduced here, is used to identify all proper nouns (replaced with NP1), punctuation was collapsed (marked as <p>) and some additional tags are marked like <SOP> for start of posting blog and <EOP> for end of posting blog to non-linguistic features of blogs. The binary classification and 3-class classification are able to handle the larger corpus. This is finely tuned model and classifiers that seem to suffer least in the scaling up the procedures. The 3-class classification does not have the percentage perception. The NB [5] performance with 4-language model on clean data needs larger training data set. In this work, all the parts of speech (POS) have not been included. *In our work this deficiency has been overcome by considering more POS*

.

3. Proposed Human Trait Model

In this work, a rule based trait modeling has been designed to identify the trait of blog authors or any text submitted by an author.

3.1 Hypothesis

Identification of human trait is based on the feature vector extracted from blogs, online diaries and emails.

3.2 Feature Vector

In this paper, we considered the significant features that help to identify the author's trait. The trait result is categorized either as low, average or high or in percentage of trait. The feature vector is generated through active features only. The size of vector is ten as we have considered following ten attributes:

1. First Person Pronoun (FPP)
2. Second Person Pronoun (SPP)
3. Third Person Pronoun (TPP)
4. Positive Adjective Words (PAW)

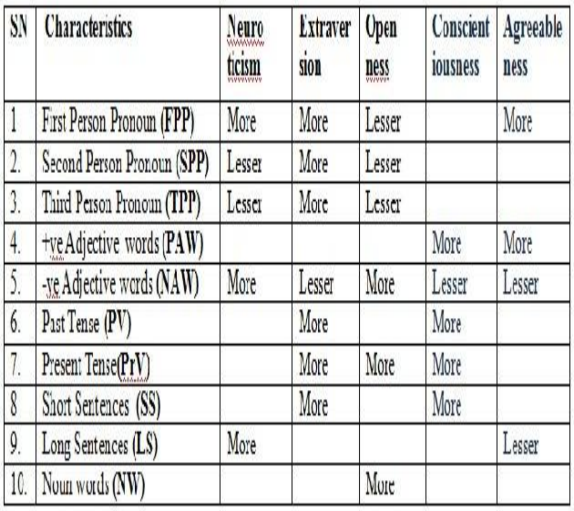
5. Negative Adjective Words(NAW)
6. Past Verbs (PV)
7. Present Verbs (PrV)
8. Short Sentences (SS)
9. Long Sentences (LS)
10. Noun Words (NW)

These attributes are taken from part of speech, definition of five-trait model [3] and from personal assessment.

3.3 Human Traits

Traits are classified as Neuroticism, Extraversion, Openness, Conscientiousness and Agreeableness. On the basis of ten attributes shown above and with reference to some previous work [4-7], the traits are defined in table1:

Table1: Trait chart



SN	Characteristics	Neuroticism	Extraversion	Openness	Conscientiousness	Agreeableness
1	First Person Pronoun (FPP)	More	More	Lesser		More
2.	Second Person Pronoun (SPP)	Lesser	More	Lesser		
3.	Third Person Pronoun (TPP)	Lesser	More	Lesser		
4.	+ve Adjective words (PAW)				More	More
5.	-ve Adjective words (NAW)	More	Lesser	More	Lesser	Lesser
6.	Past Tense (PV)		More		More	
7.	Present Tense(PrV)		More	More	More	
8	Short Sentences (SS)		More		More	
9.	Long Sentences (LS)	More				Lesser
10.	Noun words (NW)			More		

4. Methodologies

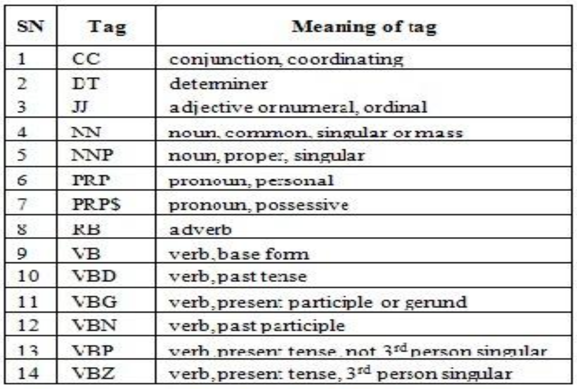
The present work is accomplished in the following steps:

(i) Tagging of the text under study using POS tagger.
(ii) Classification of Text based on defined attributes.
(iii) Generation of Feature Vector Matrix.
(iv) Designing FIS Rules for Identifying the Human Trait

4.1 Tagging of the text under study using POS tagger

The first step is to pass the input text through any tagger. In the current work, POS tagger [4] is used. Some of the tags of POS tagger and their meaning are:

Table 2: Some tags of POS Tagger and their meaning



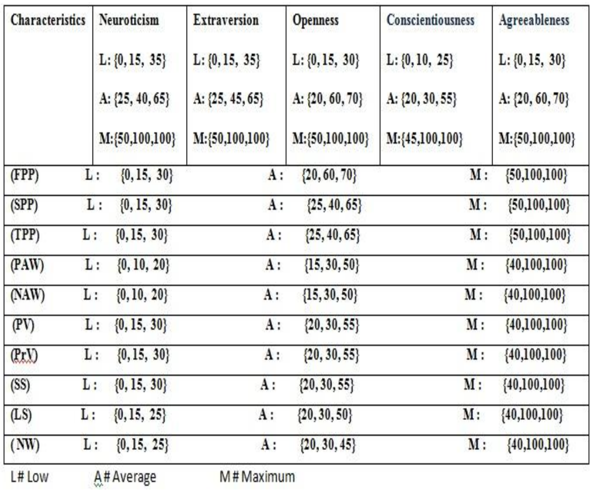
SN	Tag	Meaning of tag
1	CC	conjunction, coordinating
2	DT	determiner
3	JJ	adjective or numeral, ordinal
4	NN	noun, common, singular or mass
5	NNP	noun, proper, singular
6	PRP	pronoun, personal
7	PRP$	pronoun, possessive
8	RB	adverb
9	VB	verb, base form
10	VBD	verb, past tense
11	VBG	verb, present participle or gerund
12	VBN	verb, past participle
13	VBP	verb, present tense, not 3rd person singular
14	VBZ	verb, present tense, 3rd person singular

We used a DB table named 'words' containing all PAW and NAW. This table is created in Mysql PHP and updated every time when a new word arrives in the text. The positivity and negativity are classified on personal assessment.

4.2 Classification of Text based on defined attributes

The input text is classified on ten attributes. Each attribute and trait is categorized in three classes-less, average and maximum. The values and range for less, average and maximum are analyzed and collected on the basis of several defined texts and personal assessment. These are depicted in Table 3.

Table 3: Values of different attributes and traits



Characteristics	Neuroticism
	L: {0,15, 35}
	A: {25, 40, 65}
	M:{50,100,100}
		Extraversion
		L: {0,15, 35}
		A: {25, 45, 65}
		M:{50,100,100}
			Openness
			L: {0,15, 30}
			A: {20, 60, 70}
			M:{50,100,100}
				Conscientiousness
				L: {0,10, 25}
				A: {20, 30, 55}
				M:{45,100,100}
					Agreeableness
					L: {0,15, 30}
					A: {20, 60, 70}
					M:{50,100,100}

(FPP)	L : {0,15, 30}		A : {20, 60, 70}		M : {50,100,100}
(SPP)	L : {0, 15, 30}		A : {25, 40, 65}		M : {50,100,100}
(TPP)	L : {0,15, 30}		A : {25, 40, 65}		M : {50,100,100}
(PAW)	L : {0, 10, 20}		A : {15, 30, 50}		M : {40,100,100}
(NAW)	L : {0, 10, 20}		A : {15, 30, 50}		M : {40,100,100}
(PV)	L : {0, 15, 30}		A : {20, 30, 55}		M : {40,100,100}
(PrV)	L : {0, 15, 30}		A : {20, 30, 55}		M : {40,100,100}
(SS)	L : {0, 15, 30}		A : {20, 30, 55}		M : {40,100,100}
(LS)	L : {0, 15, 25}		A : {20, 30, 50}		M : {40,100,100}
(NW)	L : {0, 15, 25}		A : {20, 30, 45}		M : {40,100,100}

L# Low A# Average M# Maximum

4.3 Generation of Feature Vector Matrix

In this work, the size of FVM is ten. The attribute with no value is not included and has no significance in FVM. Each column of FVM is generated with its associate domain. We have calculated the participation of each attribute-PAW and NAW are calculated from total number of adjectives, FPP is calculated through total number of pronouns while SS, LS and NW are calculated through whole text. The size of short sentences (SS) is limited to ten words and long sentences (LS) is greater than ten words.

4.4 Designing FIS Rules for Identifying the Human Trait

The FVM is implemented through FIS rules designed for MATLAB7.0. The rules for Neuroticism, Extraversion, Openness, Conscientiousness and Agreeableness based on attributes defined in section 3.2 are:

4.4.1 FIS rules for Neuroticism

If (FPP is MORE) and (NAW is MORE) and (SPP is LESS) and (TPP is LESS) and (LS is MORE) then (NEUROTICISM is MORE)

If (FPP is AVG) and (NAW is MORE) and (SPP is LESS) and (TPP is LESS) and (LS is MORE) then (NEUROTICISM is AVG)

If (FPP is AVG) and (NAW is AVG) and (SPP is LESS) and (TPP is LESS) and (LS is not MORE) then (NEUROTICISM is AVG)

If (FPP is MORE) and (NAW is MORE) then (NEUROTICISM is MORE)

4.4.2 FIS rules for Extraversion

If (FPP is AVG) and (NAW is LESS) and (SPP is AVG) and (TPP is AVG) and (PV is AVG) and (PrV is AVG) and (SS is AVG) then (EXTRAVERSION is AVG)

If (FPP is LESS) and (NAW is LESS) and (SPP is LESS) and (TPP is LESS) and (PV is LESS) and (PrV is LESS) and (SS is LESS) then (EXTRAVERSION is LESS)

If (FPP is MORE) and (NAW is LESS) and (SPP is MORE) and (TPP is MORE) and (PV is MORE) and (PrV is MORE) and (SS is MORE) then (EXTRAVERSION is MORE)

4.4.3 FIS rules for Openness

If (FPP is LESS) and (NAW is MORE) and (SPP is LESS) and (PrV is MORE) and (NW is MORE) and (TPP is LESS) then (OPENNESS is MORE)

If (FPP is LESS) and (NAW is AVG) and (SPP is LESS) and (PrV is AVG) and (NW is AVG) and (TPP is LESS) then (OPENNESS is AVG)

If (FPP is LESS) and (NAW is AVG) and (SPP is LESS) and (PrV is LESS) and (NW is not LESS) and (TPP is LESS) then (OPENNESS is LESS)

4.4.4 FIS rules for Conscientiousness

If (PAW is MORE) and (NAW is LESS) and (PV is MORE) and (PrV is MORE) and (SS is MORE) then (CONSCIENTIOUSNESS is MORE)

If (PAW is AVG) and (NAW is LESS) and (PV is AVG) and (PrV is LESS) and (SS is AVG) then (CONSCIENTIOUSNESS is AVG)

If (PAW is LESS) and (NAW is LESS) and (PV is LESS) and (PrV is LESS) and (SS is LESS) then (CONSCIENTIOUSNESS is LESS)

4.4.5 FIS rules for Agreeableness

If (FPP is MORE) and (PAW is MORE) and (NAW is LESS) and (LS is LESS) then (AGREEABLNESS is MORE)

If (FPP is AVG) and (PAW is AVG) and (NAW is LESS) and (LS is LESS) then (AGREEABLNESS is AVG)

If (FPP is LESS) and (PAW is LESS) and (NAW is LESS) and (LS is LESS) then (AGREEABLNESS is LESS)

5. Implementation and results

The current work can be explained through the following diagram (Figure 1). The figure also shows the step wise method from left to right.

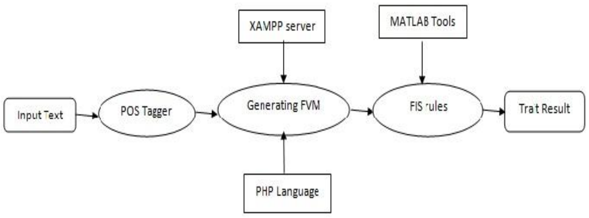


Figure1: Implementation setup through DFD

5.1 Tagging of the text

The data may be any blog, online diaries or email. One sample of author's blog is:

"Indian Team has won the Cricket World cup 2011. But we are unhappy due to inconsistent performance of team members."

Following result is found on passing this sentence to POS tagger:

Indian/NNP Team/NN has/VBZ won/VBD the/DT Cricket/NN World/NN cup/NN 2011./NN But/CC we/PRP are/VBP unhappy/JJ due/JJ to/TO inconsistent/JJ performance/NN of/IN team/NN members./NNS

5.2 Classification of Text

The attributes are observed on the text with total words twenty and their values are counted as:

FPP = 1 (100%)	NAW = 2 (67%)
PAW = 1 (33%)	SPP = 0
TPP = 0	SS = 0
LS = 1 (5%)	PV = 1 (33%)
PrV = 2 (67%)	NW = 7 (35%)

The numbers are verified and tested with the result obtained from POS tagger.

The adjectives are classified as positive or negative. We used a DB table 'words' for this purpose which is updated for each new word. For instance, 'unhappy' and 'inconsistent' as NAW while 'due' as PAW and are added in the DB table.

5.3 Generation of FVM

In the sample case, the feature vector of size ten and its values are:

FPP	NAW	PAW	LS	PV	PrV	NW
1.00	.67	.33	.05	.33	.67	.35

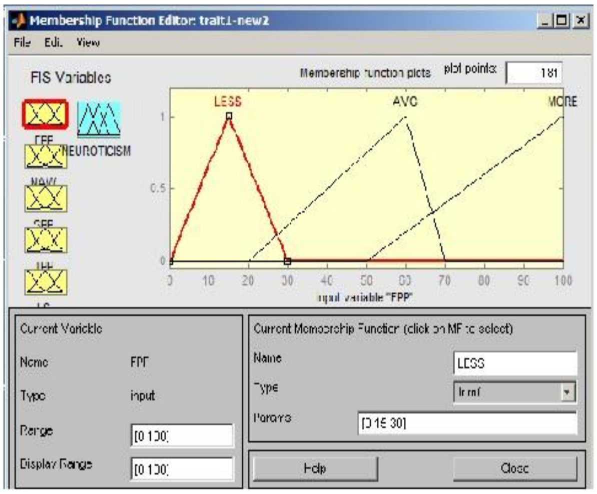
Alternatively, The Feature Vector Matrix (FVM) is

1: 1.00 **2**: 0.67 **3**: 0.33 **7**: 0.05 **8**: 0.33 **9**: 0.67 **10**: 0.35 :eq[1]

5.4 Results according to FIS Rules

The FVM shows that the maximum attributes falls in 'Neuroticism' category. So the FVM should be passed through FIS rules written for 'Neuroticism'.

We have implemented our work in MATLAB 7.0. The FIS variables are as par the Table 1 and Table 2. Some of FIS reports are

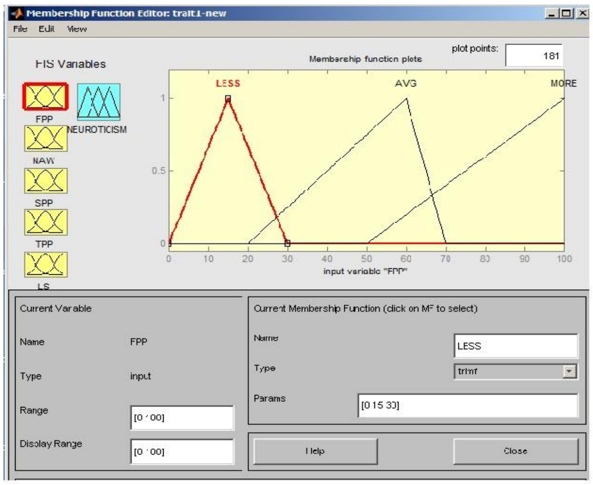


plot points: 181

Current Variable
Name FPP
Type input
Range [0 100]
Display Range [0 100]

Current Membership Function (click on MF to select)
Name LESS
Type trimf
Params [0 15 30]

Figure 2: FIS graph for FPP for Neuroticism



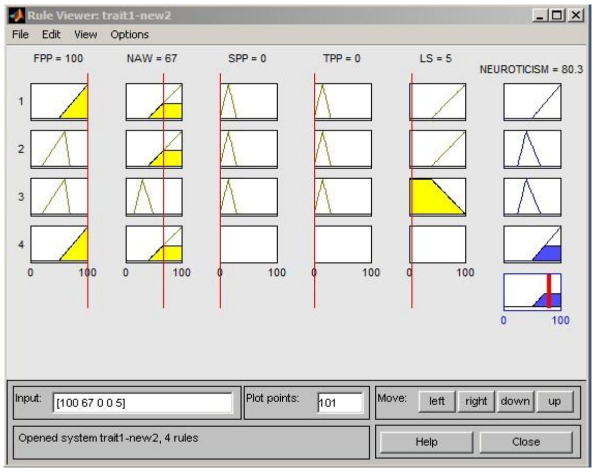
plot points: 181

Current Variable
Name FPP
Type input
Range [0 100]
Display Range [0 100]

Current Membership Function (click on MF to select)
Name LESS
Type trimf
Params [0 15 30]

Figure3: FIS graph for NAW for Neuroticism

The output for the given FVM is.



FPP = 100
NAW = 67
SPP = 0
TPP = 0
LS = 5
NEUROTICISM = 80.3

Input: [100 67 0 0 5]
Plot points: 101

Opened system trait1-new2, 4 rules

Figure 4: Result w.r.t. FVM of eq[1]

The output says the blogger is 80.3% Neuroticism.

6. Conclusion

This paper studied the problem of trait identification. Although there have been several existing papers [3, 5] studying the problem, our work gives the result in different perception. If the same sample written in section 5.1 is analyzed through the earlier study [5], it gives the text belongs to a highly neurotic author while our work gives the percentage of degree of neuroticism by using a set of FIS rules. The result obtained by using our methodology level of human trait is found useful in relative comparison of two authors with similar traits. In this work, we proposed a new class of attributes including few parts of speech and some general purpose attributes. A large number of texts and real-life blogs are tested through this model and yields in improved and much accurate result. The addition of PAW and NAW in classification improves the accuracy because the previous studies [3, 5] show that neurotic persons generally use more NAW in their texts while conscientious persons use more PAW. In the same context, the number of NW in any text is an important attribute. This paper also considered NW. In addition to other features, the attributes short sentences (SS) and long sentences (LS) are also enhancing the final outcome of trait identification. The FVM is analyzed through FIS and then implemented in MATLAB 7.0. The specific result may help in comparing the behavior of the authors can be used in various applications.

7. References

[1] Carlo Strapparava and Rada Mihalcea, "Learning to identify Emotions in Text', in *Proceedings of SAC'March 2008, ACM,* Brazil page 16-20

[2] Arjun Mukherjee and Bing Liu, "Improving Gender Classification of Blog Authors", in the *Proceedings of 2010 Conference on Empirical Methods in Natural Language Processing* page 207-217, MIT USA.

[3] Haytham Mohtasseb and Amr Ahmed, "More Blogging Feature for Author Identification", in *ACM* 2007.

[4] J. Oberlander and S. Nowson, " Whose thumb is it anyway? Classifying author personality from weblog text", in *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics and 21st International Conference on Computational Linguistics,* Sydney, Australia, 2006.

[5] Scott Nowson and Jon Oberlander, "Identifying More Bloggers", in *ICWSM* 2007 USA.

[6] J.M. Dewaele and A. Furnham, "Extraversion: The unloved variable in applied linguistic research", *Language Learning*, 49:509–544, 1999.

[7] S. Argamon, S. Dhawle, M. Koppel, and J. W.Pennebaker, "Lexical predictors of personality type", in *Proceedings of the 2005 Joint Annual Meeting of the Interface and the Classification Society of North America,* 2005.

[8] J. W. Pennebaker and L. King, "Linguistic styles: Language use as an individual difference", *Journal of Personality and Social Psychology*, 77:1296–1312, 1999.

[9] K. Scherer, "Personality markers in speech", in *K. R.Scherer and H. Giles, editors, Social Markers in Speech,* pages 147–209. Cambridge University Press, Cambridge, 1979.

[10] Yla R. Tausczik and James W. Pennebaker, "The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods" published in *"Language Style Matching Predicts Relationship Initiation and Stability"* Psychological Science January 1, 2011 22: 39-44.

[11] A. Gliozzo and C. Strapparava, "Domains kernels for text categorization", in *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005),* Ann Arbor, June 2005.

[12] A. Abbasi and H. Chen, "Applying authorship analysis to extremist-group web forum messages", *IEEE INTELLIGENT SYSTEMS,* pages 67–75, 2005.

[13] A. Abbasi and H. Chen, "Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace", *ACM Transaction Information Systems*, 26(2):1–29, 2008.

[14] J. W. Pennebaker and L. A. King, "Linguistic styles: language use as an individual difference", *Journal of personality and social psychology,* 77(6):1296–1312, Dec 1999.