iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** VI **Month of publication:** June 2021

DOI: https://doi.org/10.22214/ijraset.2021.35766

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VI Jun 2021- Available at www.ijraset.com

An Efficient Soft Computing Approach for Text Identification using Artificial Intelligence Model

Shashi Bhushan
Assisant Professor (Department of Computer Science and Engg.) IIMT College Of Engg Gr Noida UP

Abstract: ***This paper presents an enhanced system in the field of text identification using Soft computing techniques. The model designed in this work analyzes the blogs or input text and classifies the personality into five major categories; Neuroticism, Extraversion, Openness, Conscientiousness and Agreeableness. The blog or text is first passed through POS tagger then a feature vector matrix is generated according to the attributes of the personality chart. Each column of FVM is calculated in its domain that improves the final result of personality identification. The result of the proposed model is improvement over similar work by other researchers [1, 2, 3].***

Keywords: Soft Computing, Fuzzy Computing, Blog, POS Tagger, Feature Vector Matrix, Fuzzy Inference System, Fuzzy system etc.

I. INTRODUCTION

Today the most popular method to share thoughts, feelings and communicate with other people is individual blog or online diary. Some blogs focus on a fixed 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 this paper a new proposed model is completely based on Soft computing methodology chiefly using fuzzy system.

A. Soft Computing

The word Soft Computing (SC) refers to a group of computing techniques that consist of four different parts *viz.* fuzzy logic, evolutionary computation, neural networks and probabilistic reasoning. The theory of Soft Computing was introduced by L.A. Zadeh - the father of Fuzzy logic. Soft Computing is a new multidisciplinary field, to construct new generation of Artificial Intelligence, known as Computational Intelligence.

B. Fuzzy Computing

In real time most of the thing are based on fuzzy knowledge, that is, knowledge which is vague, imprecise, uncertain, ambiguous, inexact, or probabilistic in nature.[14]

C. Fuzzy Systems

Fuzzy Systems is a combination of Fuzzy Logic and Fuzzy Set Theory. Any system that uses Fuzzy mathematics is based on Fuzzy system. A block diagram of Fuzzy System is shown below:

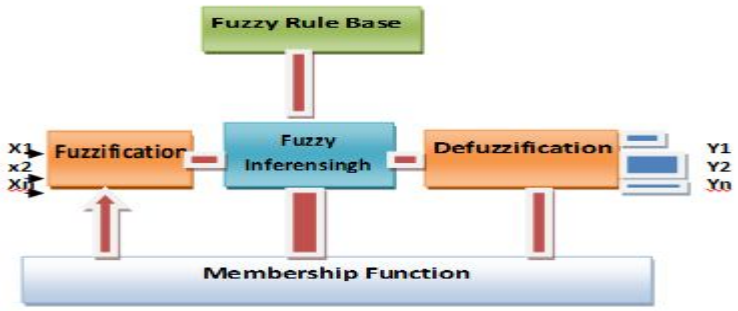


Fig. Elements of Fuzzy System

The various calculations are involved in author identification and cannot be accomplished directly but it may trap some uncertainty in human perception. On the basis of doubt, the Personality can have lower or higher values.

II. 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 personality is not adequate. The work can be enhanced by addition of more Personality. In 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 gives improved execution of the model.

A. Few Innovative Ideas for Personality Identification

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

III. PROPOSED HUMAN PERSONALITY MODEL

In proposed system, a rule based personality modeling has been designed to identify the personality of blog authors or any text submitted by an author.

A. Hypothesis

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

B. Feature Vector

In this paper, we considered the significant features that help to identify the author's personality.

The personality result is categorized either as low, average or high or in percentage of personality. 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-personality model [3] and from personal assessment.

C. *Human Personality*

Text 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 Personality is defined in table1:

Table1: Personality 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		

IV. METHODOLOGIES

The present work is accomplished in the following steps:

A. *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 ornumeral, 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 and updated very time when a new word arrives in the text. The positivity and negativity are classified on personal assessment.

B. Classification of Text Based on Defined Attributes

The input text is classified on ten attributes. Each attribute and personality 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 Personality

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

C. 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.

D. Designing FIS Rules for Identifying the Human Personality

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:

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)

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)

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) 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)

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)

V. 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.

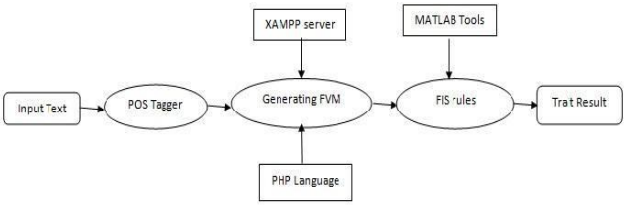


Figure1: Implementation setup through DFD

A. 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

B. 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 verification and testing of number is done on the result obtained from POS tagger.
The classification of adjective 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.

C. 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]

D. 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

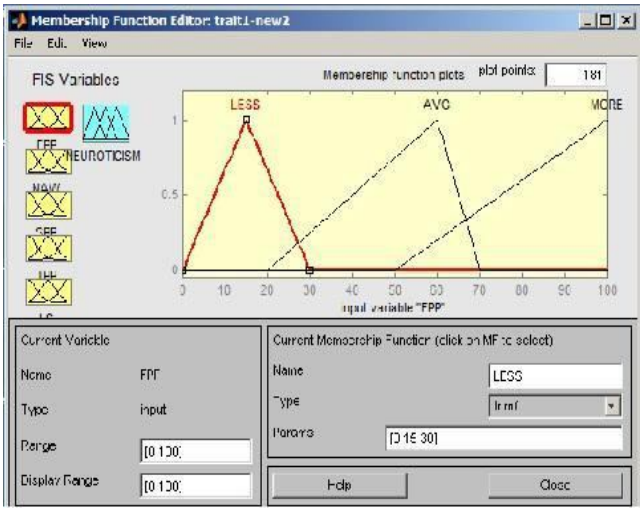


Figure 2: FIS graph for FPP for Neuroticism

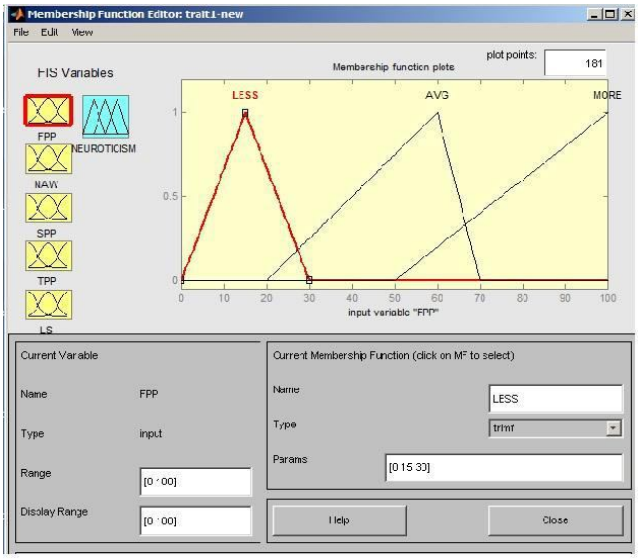


Figure3: FIS graph for NAW for Neuroticism

The output for the given FVM is.

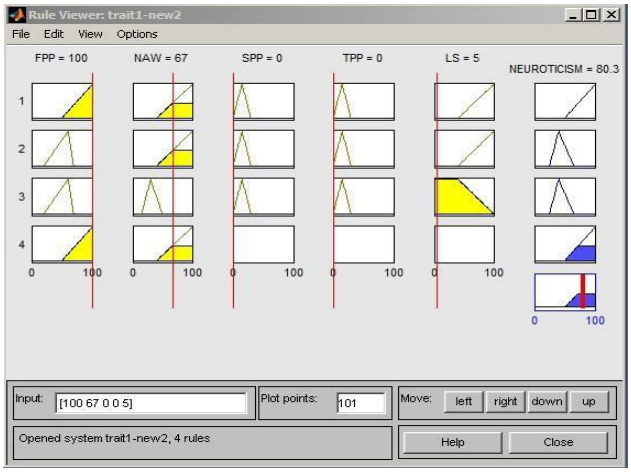


Figure 4: Result w.r.t. FVM of eq[1]

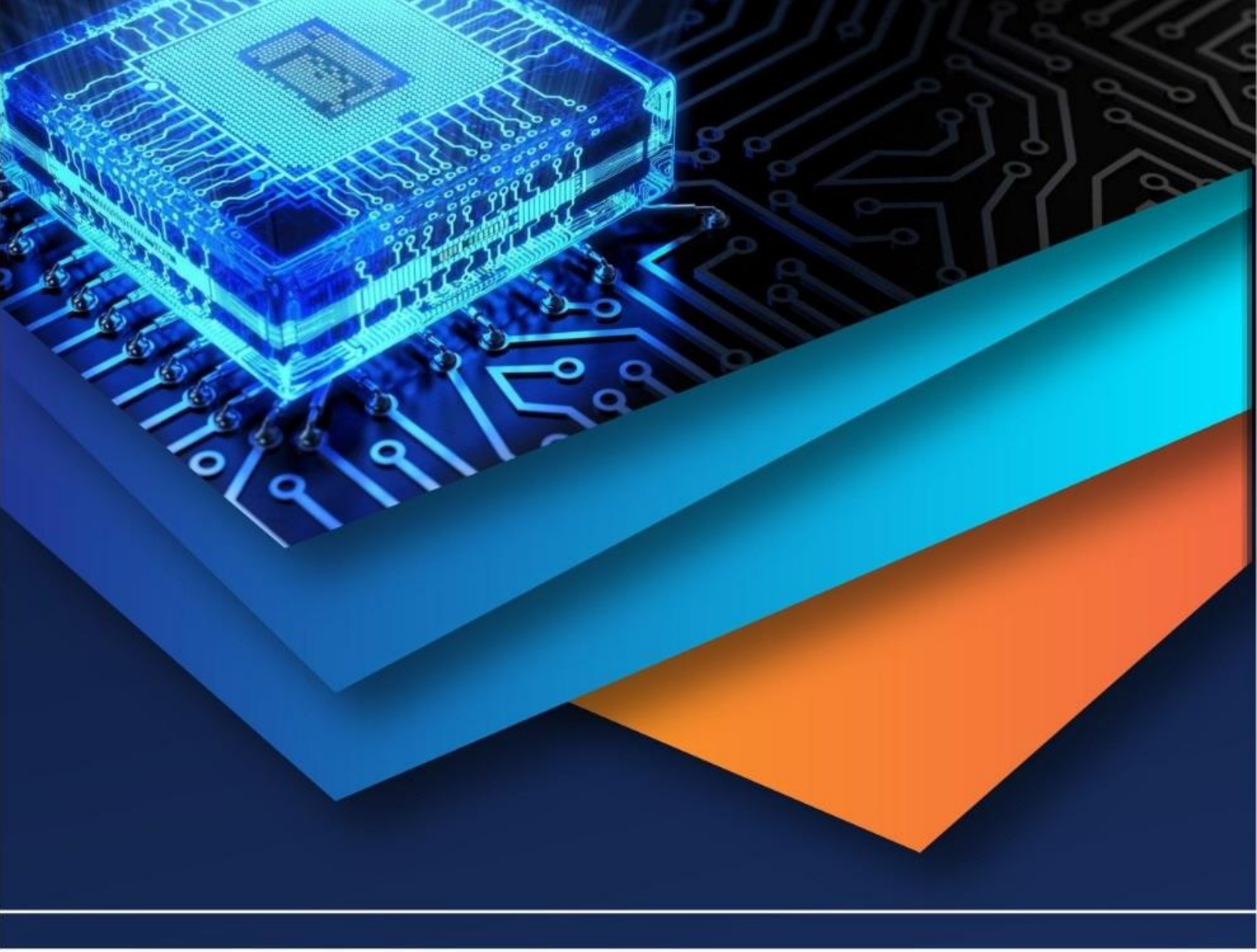
The output says the blogger is 80.3% Neuroticism.

VI. CONCLUSION

The proposed paper studied the problem of personality identification. Although there have been several existing papers [3, 5] studying the problem, our model shows 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 personality is found useful in relative comparison of two authors with similar Personality. 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 & 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 personality 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.

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 trait 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] 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.
[9] 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.
[10] 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.
[11] Abbasi and H. Chen, “Applying authorship analysis to extremist-group web forum messages”, IEEE INTELLIGENT SYSTEMS, pages 67–75, 2005.
[12] 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.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY