Machine Learning based approach for Human Trait Identification from Blog Data

Saurabh Saxena
Department of IT
Institute of Technology and Science
Mohan Nagar, Ghaziabad, India

Chandra Mani Sharma
Department of IT
Institute of Technology and Science
Mohan Nagar, Ghaziabad, India

ABSTRACT
Emotions form a major part of a person's personality. Emotional intelligence (EI) is the ability to identify, assess, and control the emotions of oneself, of others, and of groups. The written expressions reflect author's personality. Various personality traits can be determined by the analysis of the contents written by a person. This paper proposes a novel technique for human trait identification from the analysis of author's written expressions. The proposed technique is based on the concept of supervised machine learning and uses Support Vector Machine for classifying the personality of a writer. We classify the personality of a writer into five categories namely, highly extrovert, highly introvert, low introvert, low extrovert and ambivert. Experiments have been carried out on the real world blog data and results demonstrate that the proposed technique can determine the personality traits of a writer with accuracy and speed. We have also implemented a PHP based online system, which reads the contents of a blog and can automatically predict the personality of writer of the blog

General Terms
Algorithms, Text Processing, Emotion Mining, Human Behavior Analysis, Pattern Recognition.

Keywords
Machine Learning, Soft Computing, Support Vector Machine, Author Trait Identification, Online System,

1. INTRODUCTION
The English word 'emotion' is derived from the French word 'emouvoir'. This is based on the Latin "emovere", where 'e' means out and 'movere' means move. Emotion is the complex psycho physiological experience of an individual's state of mind as interacting with biochemical and environmental influences. Every person uses natural language for communicating with other persons. This communication takes place either verbal or in writing. Emotion mining is the area for finding the emotion in written text or spoken words. In every natural language there are some specified words, which show some emotion in verbal or written communication. So it is quintessential to tag those words, which show some emotion relevant as per the requirement of the application domain. Analysis of usage of these words can help in prediction of emotions. Parts of Speech tagger can play a helpful role in tagging words in a text. Emotion mining is a multi-disciplinary domain and involves the application of natural language processing (NLP), computational linguistics, and text analytics to identify and extract subjective information from source materials. Generally speaking, sentiment analysis aims at determining the attitude of a speaker or a writer with respect to some topic or the overall tonality of a document. The attitude may be individual's judgment or evaluation, affective state, or the intended emotional communication. A very limited research work has been carried out in the area of automated writer's trait identification. Poropat et al.[1] analyze and improve the Big Five theory of human trait identification. Here, a large blog corpus is used to examine the role of personality in motivation for blogging. They hypothesized human personality to have five traits- *Neuroticism, Extraversion, Openness, Agreeableness and Conscientiousness*.

Linguistic content of the blogs revealed cathartic and auto-therapeutic tendencies of high Neurotic bloggers; high level life documentation and emotion expression by high Extroverts; commentary and evolution by high Openness scorers and reports of daily life by high Conscientiousness bloggers. Highly Agreeable authors demonstrate similar strategies to those adopt in other contexts. Gill et al [2]discuss about the personality of authors on blog data. Blog is a place in Internet where a person shares his views about any entity. Blogs provide the individuals an opportunity to write freely and express themselves online in the presence of others. In this paper author discuss about the properties of different personality traits. These traits are shown in Table 1.

Table 1: Human Personality and Corresponding Emotion Reflection in Blog Writing

Trait	High Behaviours	Low Behaviours
Neuroticism	emotional instability; anxious; hostile; prone to depression	emotional stability; calm; less easily upset
Extraversion	extraverts; warmer; more assertive; action-oriented, thrill-seeking	introverts; low-key; deliberate; require less stimulation
Openness	appreciation for art and ideas; imaginative; more aware of feelings	more straightforward interests; conservative; resistant to change
Agreeableness	compassionate; cooperative; considerate; friendly	suspicious; unfriendly; wary; antagonistic; uncooperative
Conscientiousness	disciplined; dutiful; persistent; compulsive ; perfectionist	spontaneous; impulsive; less driven by desire for achievement

Mohtasseb et al.[3]present a method for authorship identification in personal blogs. Their approach makes use of linguistic features that best capture the style of users in dairy blogs. Here, authors use the feature set containing LIWC with

its psychology background and a collection of syntactic features along with Parts of Speech (POS). Abbasi et al.[4] discuss about the analysis of an author's writing style in context of web forums. Here, they use a collection of lexical, syntactical, structural, and content specific features to find out the extreme patterns of writing on web forums. Text in the web forums is similar to that in the personal blogs. Yet, in case of web forums there is some specific topic to discuss in case of web forums while in case of personal blog (diary) the author is free to express without sticking to a single topic. In their another paper, Abbasi et al.[5]present the "Writeprints" technique, which separately model the features of each individual author, instead of using one model for all the authors. They build a write print for each author using the author's key features. This approach is computationally expensive as multiple models need to be created and managed. Vel et al.[6] discuss about the authorship identification in context of email. Although they achieved relatively good results, this may not be applicable straight-forward on the blogs due to the different nature of the text in emails and blogs. Generally, email text is shorter than blog text. Email text is usually a topical dialogue between two authors, while online blog text is from the author to the public, at least the intended group. Oberlander et al.[7] report on initial results to classify the four personality traits of authors. Mukharjee et al.[8] discuss the problem of automatically classifying the gender of a blog author.

2. THE PROPOSED METHOD

We hypothesize the following five traits of blog authors:

i. **High Extrovert:** The bloggers belonging to this trait category mainly use more social words, always referencing to themselves and others. They use words with positive emotions and express with certainty while writing. This type of personality shows increased use of introducing clause-initial connectives such as then, which and what conjunctions and adjectives for writing a E-mail. While writing a blog such type of persons use more present tense verbs in communication.

ii. **Low Extrovert:** Low Extrovert bloggers use more negations exclusive, inclusive and causation words for emotion expressions. They use tentative articles. In blog writing, Low Extrovert person talks about achievements and use words relating to discrepancies.

iii. **High Introvert:** Persons belonging to this category have high scorers in monologue situations and have been found using singular and negative emotion words with more first person pronoun. Besides, they excessively talk about discrepancies, jobs and physical states. Additionally they use less outward-looking discourse and their discourse contains fewer phrases referring to others. They also use more exclusive, inclusive connectives with a greater use of multiple punctuation expressions in essay writing.

iv. **Low Introvert:** Low Introvert person always refers more to other people and use more nouns and adverbs. While writing, they have high Openness scorers and use more articles, rather longer insight words, and fewer first person sentences and causation words. In Blog writing, they use rather longer words and also express positive feelings as well as use inclusive words. They also use fewer negations and write less about the school!

v. **Ambivert:** There are mainly four categories of traits which we have discussed above. A person who does not belong to any of the aforementioned categories is placed in Ambivert category.

Our approach uses the concept of machine learning. The system is trained, tested it and then used to function with the real-world data. The proposed technique has following 5 steps:

a. Collection and Cleansing of training data.
b. Tagging the training data.
c. Generation of Feature Vector Matrix
d. Training the system using Support Vector Machine
e. Testing the system.
f. Operating the system with actual data.

2.1 Collection and Cleansing of Data

We collect the training data and remove the inconsistencies. For data collection we used the various data sources including digital dictionary Word Net, web data and e-books etc. The inconsistencies such as related to word sense disambiguation (WSD) were resolved here.

2.2 Tagging The Text

The basic assumption behind our research is that every word, in a given context, is distinguishable from the others and is the representative of a human emotion in that context. Thus we mark the word in our training data with a unique tag. The basic role of this step is to tag all the words of text.

Table 2: Different Tags and Their Meaning

Tag	Meaning of tag	Tag	Meaning
CC	conjunction, coordinating	RB	Adverb
CD	numeral, cardinal	RBR	adverb, comparative
DT	Determiner	RBS	adverb, superlative
EX	existential there	RP	Particle
FW	foreign word	TO	"to" as preposition or infinitive marker
IN	preposition or conjunction, subordinating	UH	Interjection
JJ	adjective or numeral, ordinal	VB	verb, base form
JJR	adjective, comparative	VBD	verb, past tense
JJS	adjective, superlative	VBG	verb, present participle or gerund
MD	modal auxiliary	VBN	verb, past participle
NN	noun, common, singular or mass	VBP	verb, present tense, not 3rd person, singular
NNP	noun, proper, singular	VBZ	verb, present tense, 3rd person singular
NNPS	noun, proper, plural	WDT	WH-determiner
NNS	noun, common, plural	WP	WH-pronoun
POS	genitive marker	WP$	WH-pronoun, possessive
PRP	pronoun, personal	WRB	WH-adverb
PRP$	pronoun, possessive		

2.3 Generation of Feature Vector Matrix

In this method we generate a feature vector matrix of the text. Feature Vector matrix is nothing but a method for

representing the text on the basis of its characteristics. These characteristics are predefined or we can say that under what feature we want to categorize the text. As an example in our proposed algorithm we have taken 13 features of the text e.g: Past Tense, Positive Adjective. In feature Vector Matrix columns represent the attributes of the text. There is a requirement of generation of Feature Vector Matrix because Support Vector Machine takes input in a special format.

2.4 Classification with Support Vector Machine

Support vector machines (SVMs) are a set of related supervised learning methods that analyze data and recognize patterns, used for classification and regression analysis. The standard SVM takes a set of input data and predicts, for each given input, which of two possible classes the input is a member of, which makes the SVM a non-probabilistic binary linear classifier. Support vector machine constructs a hyper plane or set of hyper planes in a high or infinite dimensional space, which can be used for classification, regression, or other tasks. For using the Support Vector Machine we give input in this in a very specific format which as follows:

<label> <index1>:<value1> <index2>:<value2>……
<index n>:<value n>

where, <label> is the target value of the training data. For classification, it should be an integer and identifies a particular class of data.

In the proposed algorithm for classifying a blog writer into a suitable trait category, we postulate two cases. In case 1, we can check for whether an author is Highly Introvert, Low Extrovert or Ambivert. In case 2, we can check for whether the author is High Extrovert, Low Introvert or Ambivert.

The classification criteria, used in the proposed algorithm, have also been illustrated in Table 3. This table takes into consideration some 10 odd check points, necessary in categorization of the traits of a blog author.

ALGORITHM

Case 1:

If (60% or more negative emotion adjectives)
  If (excessive use of first person pronouns)
    If (maximum short sentences)
      Then "highly introvert"
    Else "Ambivert"
  Else "Low Extrovert"
Else "Ambivert"

Case 2:

If (60% or more adjectives have positive emotion && maximum present tense)
  If (past tense + present tense >50%)
    If (more than 60% pronouns are third person pronoun)
      If (conjunctions>5%)
        Then "High extrovert"
      Else "Ambivert"
    Else "Ambivert"
  Else If (articles>=10%)
    If (1st person pronoun>=60%)
      Then "Low Introvert"
    Else "Ambivert"
  Else "Ambivert"
Else "Ambivert".

In Table 3, a right mark depicts that a particular condition holds true in the given case and the absence of right mark represents falsity of the condition. The classification criteria have certain key parameters and incorporate excessive and rare use of different types of pronouns, various tenses, average number of words in the sentences in the reference blog, use of conjunctions, articles, negative and positive emotions etc. For figuring out the positive and negative emotions, certain special words are tagged, in the text tagging phase, to show sense of pessimism or optimism of the author.

Table 3: Criteria for Trait Classification

Property	Highly Introvert	Highly Extrovert	Low Introvert	Low Extrovert
Excessive use of first person pronoun	✓			
Excessive use of negative emotion adjective	✓			✓
Short sentences	✓			
Excessive use of past tense		✓		
Excessive use of present tense		✓	✓	
60% or more pronouns are third person		✓		
5% or more words are conjunctions		✓		
60% or more adjectives project positive feelings		✓	✓	
10% or more words are from articles			✓	
60% or more pronouns are first person			✓	

3. EXPERIMENTAL RESULTS AND DISCUSSIONS

For testing the reliability of any machine learning based system a test data set is required. We have tested the proposed system with multiple data sources and set-ups.

For exhaustive performance evaluation of the proposed system, we created a data base having diversified essays as its contents. Moreover, the blog data available on web was also used. For ascertaining the accuracy of experimental results, we considered that blog data only for which author's traits were known and could be known with certainty. Therefore the ground truth tag values for different authors were collected and created.

To begin with, a group of 50 odd blog authors was identified. We first conducted a questionnaire and personal interview to manually classify them into five hypothesized classes of personality traits viz. *High Extrovert, Low Extrovert, Ambivert, Low Introvert* and *High Introvert.*

The consideration for manual tagging of blog authors based on there responses to the questions in the questionnaire is given in Table 4. This tagged data having ground truth values was also used to test the accuracy of the proposed classification system.

Following assumptions have been laid down in the working set of the proposed system:

i. Full Stop must not be attached to any word.
ii. It is assumed that the text to be tested does not have any typographical errors.
iii. That the text typed by a person (in blog) has at least 100 words.
iv. Entered text is punctually and grammatically correct.
v. Author uses moderated length of sentences without slang words.

After the multiple runs on seen and unseen data, the average accuracy of the proposed system has been calculated as 89.25%.

Table 4: Ground Truth Tagging of the Authors

Question	High Extrovert	Low Extrovert	Ambivert	Low Inrovert	High Introvert
You are the life of the party?	Agree	Agree	Agree	Disagree	Disagree
You enjoy being the center of attention?	Agree	Disagree	Agree	Disagree	Disagree
You are skilled in handling social situations?	Agree	Agree	Agree	Disagree	Disagree
You like to be where the action is?	Agree	Agree	Disagree	Agree	Disagree
You make new friends easily?	Agree	Agree	Disagree	Agree	Disagree
You are quiet around strangers?	Disagree	Disagree	Agree	Disagree	Agree
You don't like to draw attention to yourself?	Disagree	Agree	Agree	Agree	Agree
You don't like to party?	Disagree	Disagree	Disagree	Agree	Agree
You like to work independently?	Disagree	Agree	Agree	Agree	Agree
You often enjoy spending time by yourself?	Disagree	Disagree	Disagree	Agree	Agree

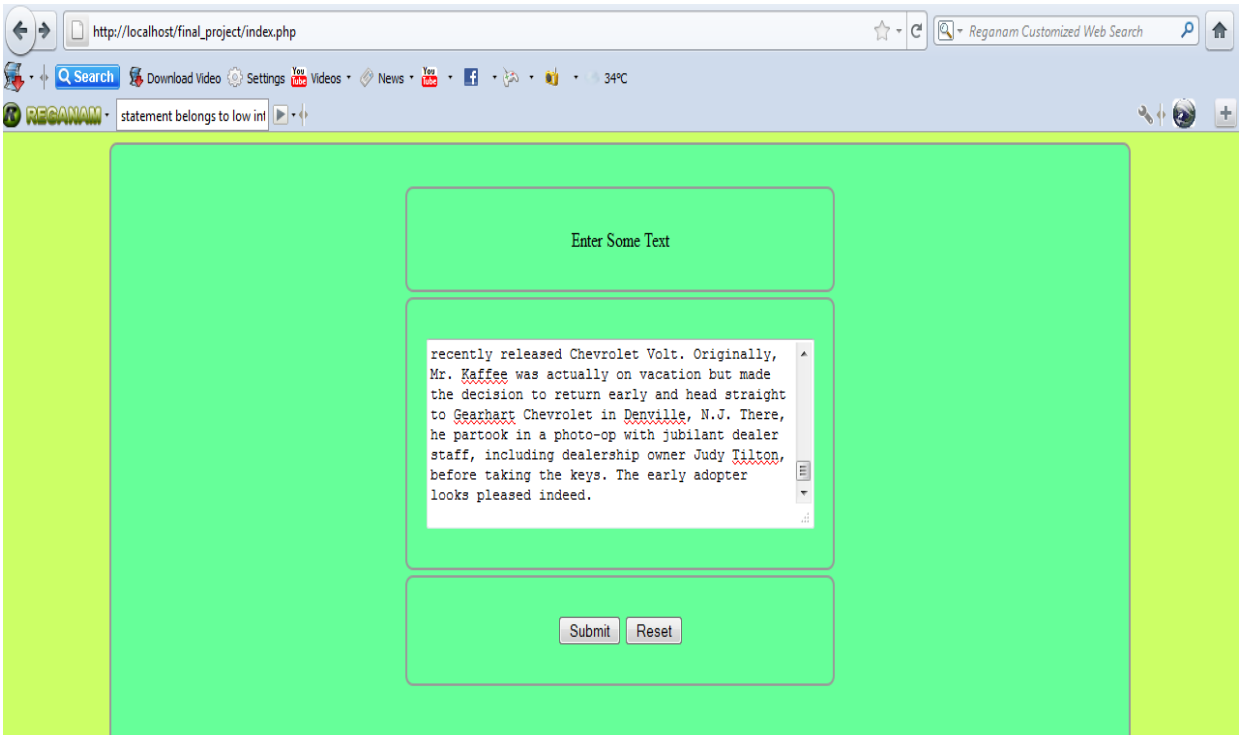

Fig. 1: User Interface for Blog Data Entry

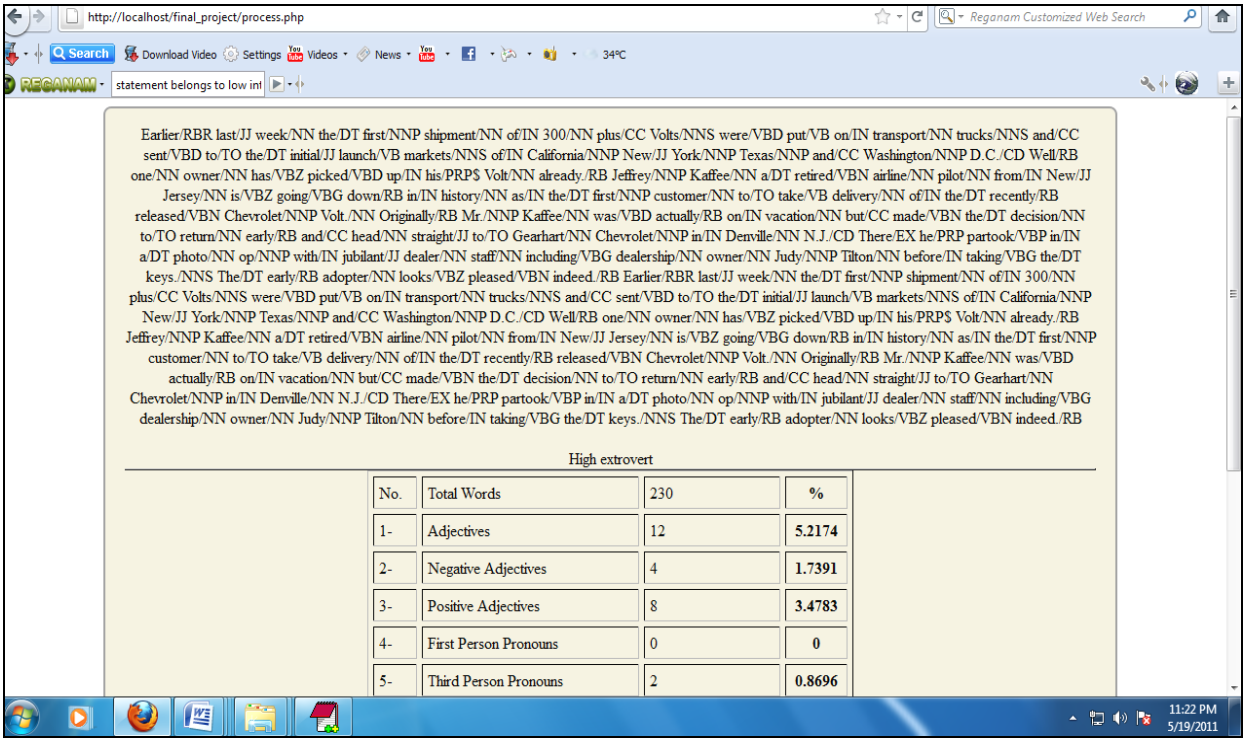

Fig. 2: Tagging of Text in Blog Data

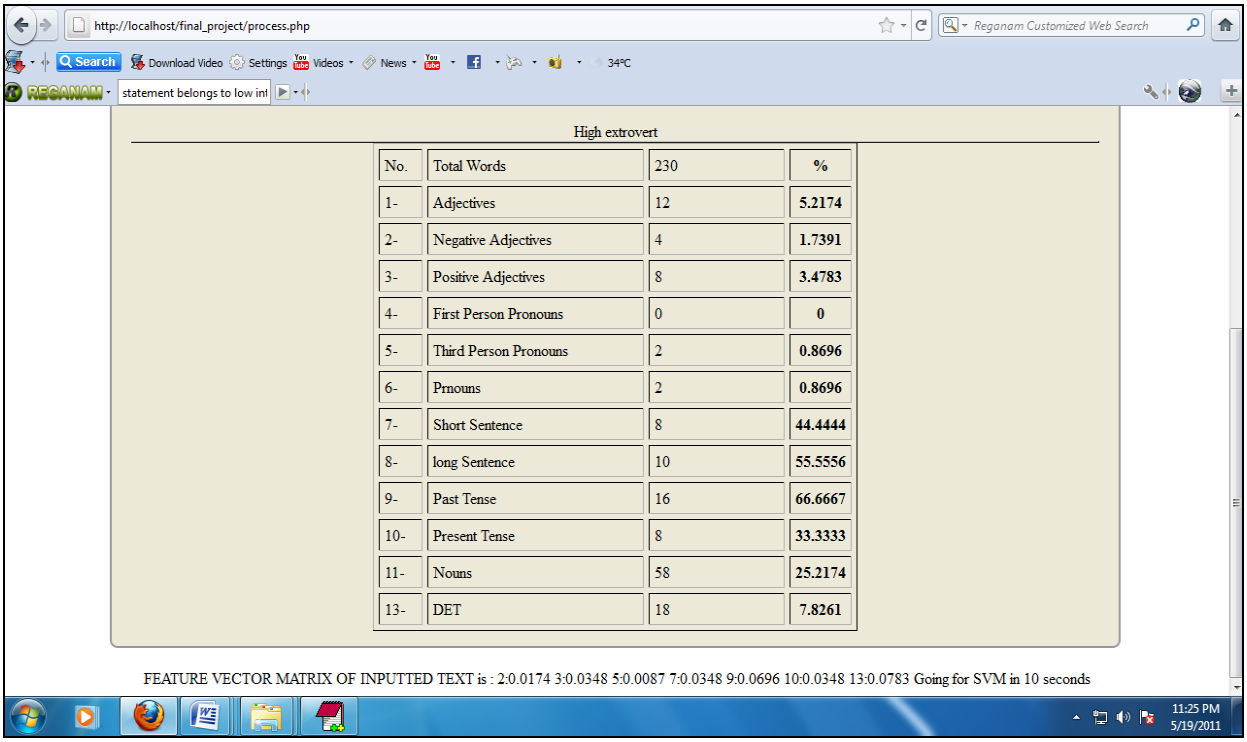

Fig. 3: Generation of Feature Vector Matrix and Analysis of Words

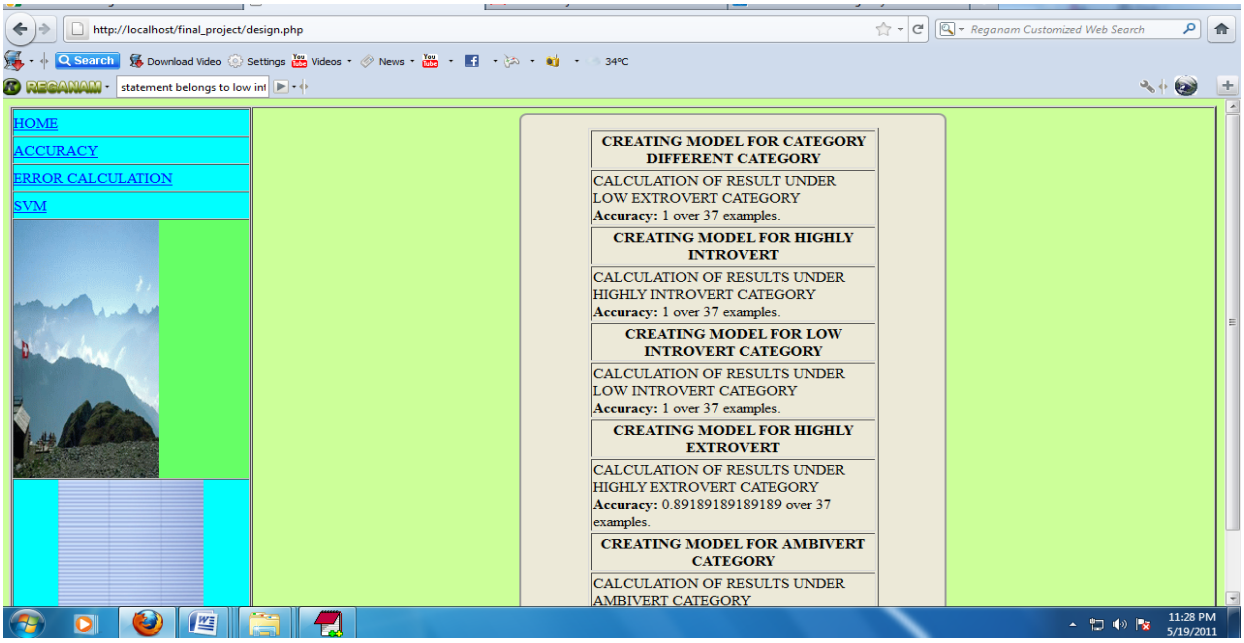

Fig.4: Classification of Traits using Support Vector Machine

The user interface of the system has been depicted in Figure 1,2,3 and 4. The text the blog is entered through the text field shown in figure 1. In the second step of processing all the words in the text are tagged such as nouns, pronouns, articles, positive/negative emotion word etc., as shown in figure 2. The third step generates FVM(Feature Vector Matrix) of inputted text along with the refined tagging. In fourth step the classification of traits takes place with the help of SVM.

4. CONCLUSIONS

This paper proposes a novel technique for author's trait identification in context of web blog data. The personality trait identification from an individual's blogging style has various useful applications such as in target-marketing, business analytics, medical, psychology etc.. We use a machine learning approach to train the system with the help of Support Vector Machine. System is trained on an tagged corpus. Effective classification criteria are hypothesized to classify a blog author into five trait categories viz. High Extrovert, High Introvert, Low Extrovert, Low Introvert and Ambivert. The online system has been implemented in PHP. Experimental results demonstrate that the system can accurately classify a blog author into a justified trait category with an average accuracy of 89.25% on unseen data.

5. ACKNOWLEDGMENTS

We are thankful to Institute of Technology and Science, Ghaziabad for supporting and encouraging the research work. We acknowledge the contribution of different blog writers in preparing the ground truth value base for training and testing the proposed system and we also thank to Shunni-Munni for incessant encouragement and motivation.

6. REFERENCES

[1] Poropat, A. E. "A meta-analysis of the five-factor model of personality and academic performance. Psychological Bulletin," 135(2), 322–338, 2009

[2] Alastair J. Gill, S. Nowson and J. Oberlander "What are they Blogging About? Personality, Topic and Motivation in Blogs," Association for the Advancement Intelligence 2009.

[3] Haytham Mohtasseb and Amr Ahmed (2009): "More Blogging Features for Authorship Identification". In Proceedings of International Conference on Knowledge Discovery (ICKD'09), Philippines.

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

[5] 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.

[6] O. de Vel, A. Anderson, M. Corney and G. Mohay. "Mining email content for author identification forensics". ACM SIGMOD record, 30(4):55-64, 2001.

[7] 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.

[8] Arjun Mukharjee, Bing Liu "Improving Gender Classification of Blog Authors" Proceedings of the 2010 conference on Empirical Methods in natural Language Processing, pages 207-217 MIT. Massachusetts, USA 9-11 October 2010.