I.J. Information Technology and Computer Science, **2012, 10, 37-43**
Published Online September 2012 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2012.10.04

A Text Analysis Based Seamless Framework for Predicting Human Personality Traits from Social Networking Sites

Ramya Sharada K
ITC Infotech, Bangalore, Karnataka, India
aditi.ramyaaa@gmail.com

Arti Arya
PESIT, Bangalore South Campus, Karnataka, India
artiarya@pes.edu

Ragini S
ITC Infotech, Bangalore, Karnataka, India
raginis1012@gmail.com

Harish Kumar
PESIT, Bangalore South Campus, Karnataka, India
harry.softer@gmail.com

Abinaya G
Wipro, Bangalore, Karnataka, India
abinaya.g001@gmail.com

Abstract— Predicting human behavior based on the usage of text on social networking sites can be a challenging area of interest to a particular community. Text mining being a major interest in Data Mining has vast applications in various fields. Clients can assess an individual's behavior using the proposed framework that is based on person's textual interaction with other people. In this paper, a framework is proposed for predicting human behavior in three phases- Text Extraction, Text cleaning and Text Analysis. For cleaning text, all the stop words have been removed and then the text is utilized for further processing. Then, the terms from the text are clustered based on semantic similarity and then gets associated with respective physiological parameters that identify a human behavior. This application is best suited for the fields of Criminal Sciences, Medical Sciences, Human Resource Department and Political Science and even for Matrimonial purposes. The proposed framework is applied on some famous world known celebrities and the results are quite encouraging.

Index Terms— Clustering, Classification, Social Networking Sites, Text Summarization, Text Analysis

I. Introduction

In current scenario, e-socializing is a rage, people like to update each and every information regarding their day-to-day life on web. They share very minute details through social networking sites. In general, if people are adopting this way of socializing, then analyzing this data, that is openly available on web can be of great use in multiple situations. For example, even criminals have such accounts on social networking sites. Analyzing their accounts on web can be of great help to investigating agencies.

Analysis of such text involves concepts of text mining [1] also. Text mining is used to convert text in structured format and explore hidden interesting patterns occurring in text [8].

There are many algorithms available in literature [1,8] for analyzing text. To the best of our knowledge no such approach is available for analyzing human behavior on the basis of his usage of text on Internet. Though literature is available wherein using text data from social networking sites, the person's gender is identified as male or female. But the characteristics of the person are not identified.

Also, "Web 3.0" or "Semantic web" is the third generation web where Internet experts believe that it will provide a unique ID for each individual profile based on the user's browsing history [5]. Web 3.0 will use this profile to tailor the browsing experience to each individual. Also, it will be able to analyze complex search queries and organize the search results in an easily understandable form. So, instead of making multiple searches, the user will be able to query a complex question. That is Web3.0 will be artificially intelligent. That means that if two different people each performed an Internet search with the same keywords using the same service, they'd receive different results determined by their individual profiles. However, this system is in its development stage [5]. But in the proposed framework, the text is extracted from the social networking site, then cleaned by removing stop words and extracting stem words. Further that text data is put forward for analysis wherein the behavior traits of the user are identified.

The rest of the paper is organized as follows: Section II gives the motivation behind the proposed framework. Section III highlights the details of the proposed framework, Section IV provides the description about the experimental results conducted on the extracted text from the web and Section V concludes the proposed framework and gives details about future course of plan.

II. Motivation for the Proposed System

Before moving to the motivational factor behind the proposed system, some basic concepts of text mining are discussed below.

Text Preprocessing: In the extracted text from the web, various terms are present in the text data that are irrelevant and inconsequential. Such terms do not provide any important information for further analysis. These terms include stopwords such as “is”, “the”, “which”, “are”, “am” etc. The list of these terms is quite long. In text cleaning, these stopwords are removed as text preprocessing step, before feeding text as input to the proposed framework.

Text Analysis: The cleaned text is given as input to the proposed system, where text is analyzed on the basis of the built in psychological parameters in the system.

Mining text to gain knowledge about a person's behavior can be an area of interest to a particular community. Now a day, usage of sites for exchanging information regarding self or others has become very common. For certain situations, analyzing this text can be of great use to understand one's behavior. However, the application areas of this framework are discussed further in Section 3.

At earlier times the modes of communication were restricted to letters and phone calls. These modes had their own drawbacks stated as with letters, time factor and change of address which may lead to the letters been undelivered and as with telephone, though it was the only means of real-time communication, it was expensive and also exposed to noise. Recently, these shortcomings are overcome by the usage of networking sites. The usage of networking sites is all the rage as of today; people are exposed to these sites and are more excited in imparting their day-to-day activities on these sites. This motivated us to play with text.

Also, there are many existing methods for predicting human behavior. This analysis may be based on a person’s handwriting or on his date of birth or his/her gestures and so on. Handwriting Analysis is done based on the keystrokes of the alphabets of a person [15], Gesture Analysis is done based on the body language or how a person uses his body to convey different messages to others without speaking. Numerology is more related to the date of birth of an individual and so on. To increase the accuracy of the results obtained from the above methods, text can also be considered as an important factor.

Thus this framework is built to interpret the personality of a person by extracting the chat cum textual interaction from these sites of an individual’s account. This framework finds its scope in the fields of Criminal Sciences, Medical Sciences and in Human Resource Department of an Organization to name a few.

A lot of research has been done in the area of text mining, information retrieval, machine learning, natural language processing and understanding [3, 8]. In [10], authors have explained a framework for human behavior understanding and prediction in three stages. The three stages are Data Management, New Knowledge Generation, Service Exposure and Control. For detailed overview of the framework, [10] can be referred.

In [11], authors have suggested an approach for predicting human gender from their chats. They have used term and style based classification for the purpose. They have also explored the effect of gender on writing style and considered many stylistic features such as sentence lengths, word usage, word lengths etc. for classification. In [12], authors have explored the feasibility of predicting user and message attributes using text in messaging services.

In [13] authors have proposed a framework that can automatically identify the gender of the users. In [14], Mats Dahllof has suggested an automatic classification approach for Swedish Politicians based on their speeches that they have delivered at various places. The author has used support vector machines for the purpose of classification.

III. Proposed System

In the proposed framework, the major area of thrust is text (chat messages) rather than the individual’s browsing history as in Web 3.0. This framework has three major phases:

1. Text Extraction
2. Text Preprocessing and Summarization
3. Text Analysis

For simplicity, the social networking site, Twitter, is considered for textual extraction. This extracted text is subjected to cleaning, summarization and analysis. Text cleaning involves removal of stop words. The stopwords are those words like prepositions, conjunctions, articles, and punctuations. Stopwords do not provide relevant information required for analysis, so can be eliminated without any loss of information. Further, replacing semantically similar terms with an appropriate term summarizes the cleaned text. This is followed by clustering of semantically similar terms of the text by using Wordnet. The clustering of terms is based on agglomerative concept [2], wherein each term is considered as an individual cluster and based on similarity between each term, grouping of the terms happen as shown in fig 1. For this phase, WordNet [7] is used, which helps in finding semantically similar words.

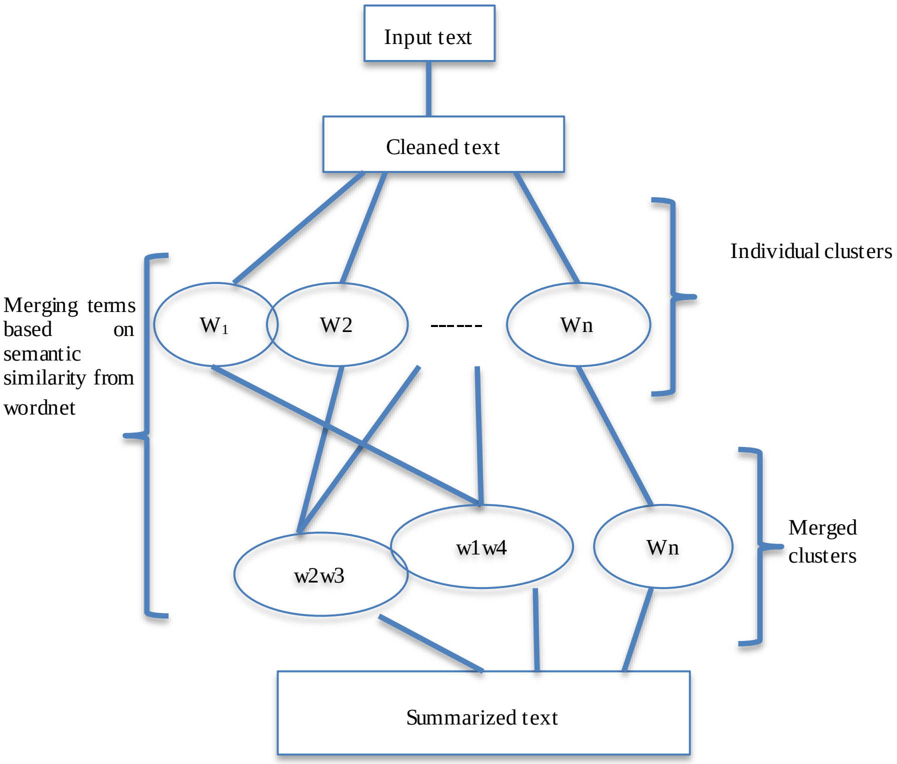


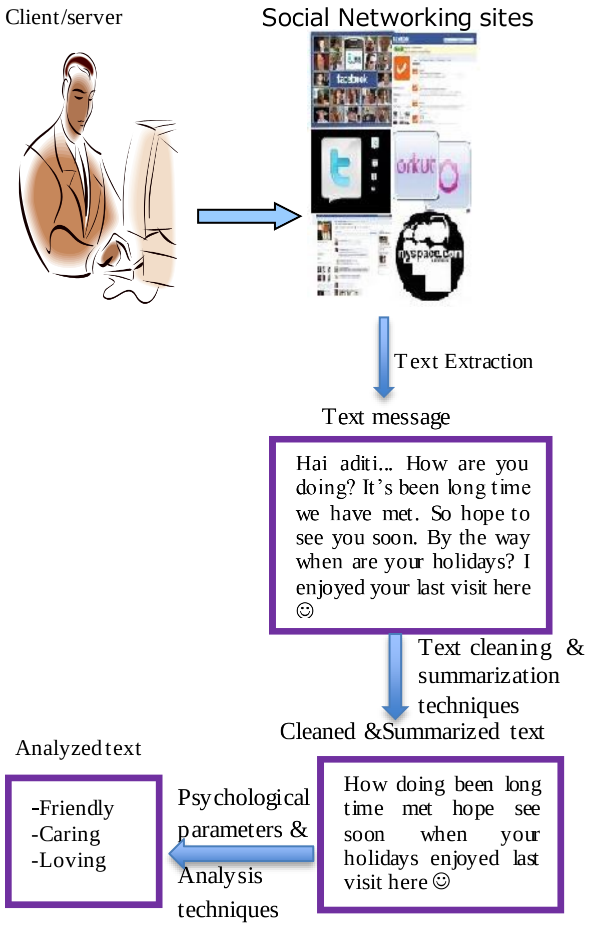
Fig. 1: Merging of term clusters based on semantic similarity

WordNet [4,6,7] is a vast lexical database developed by George Miller at Princeton University. In WordNet, different parts of Speech are clustered into various sets of synonyms known as "synsets". WordNet can be considered as a huge dictionary wherein it groups words based on their meaning.

```
Proposed Algorithm
Step 1: get text from Gui using t.get()
Step 2: Parse the smiley
Step 3: Remove the punctuations
Step4: Get stop words using "import stop words" and remove
Step 5: Finding Semantic Similarity between words using wordnet
Step 5a: get synset of w1
Step 5b: get synset of w2
Step 5c: compare (synsets (w1, w2)) if (val > Th) Tag True for those words.
Step 6 : Replace words
         for index in range (0,len(syn(w1)))
              for j in range (1,len(syn(w2)))
                 if Tag = true
                    replace w2 with w1
```

By the end of this phase, semantically similar words form a single cluster, else each word remains as a single cluster. This summarized text is an absolute input, applicable for the next phase i.e., analysis. Analysis of this text helps in identifying the behavior of the person for whom the text has been extracted. Human behavior is analyzed based on the certain important psychological parameters [16] considered in this framework. Thus the text plays the foremost role to analyze the Human Behavior.

The system architecture is as shown in fig 2, here the client is one who is interested in knowing about one's behavior, he might either be an authorized head of criminology, a person who is involved in matrimony etc., all that must be done to extract an allied person's behavior is to follow him in his social networking site, here Twitter.



Hai aditi... How are you doing? It's been long time we have met. So hope to see you soon. By the way when are your holidays? I enjoyed your last visit here ☺

How doing been long time met hope see soon when your holidays enjoyed last visit here ☺

-Friendly
-Caring
-Loving

Fig. 2: System Architecture of the proposed system

The following algorithm for analyzing the text is used for handling the text that is extracted from the Twitter and then preprocessed.

```
Text analysis Algorithm:
Step 1: Define the list of psychological parameters
using dictionaries.
Step 2 : for index from 0 to len(list(param))
            for words in sum
            define ptr j pointing to dictionary
            for each word in dictionary of particular
param
        do
            get synonymof words in above
list
             if (syn(words)in sum = syn(words) in
dictionary)
              print param
```

The flow of the application is shown by fig 3.

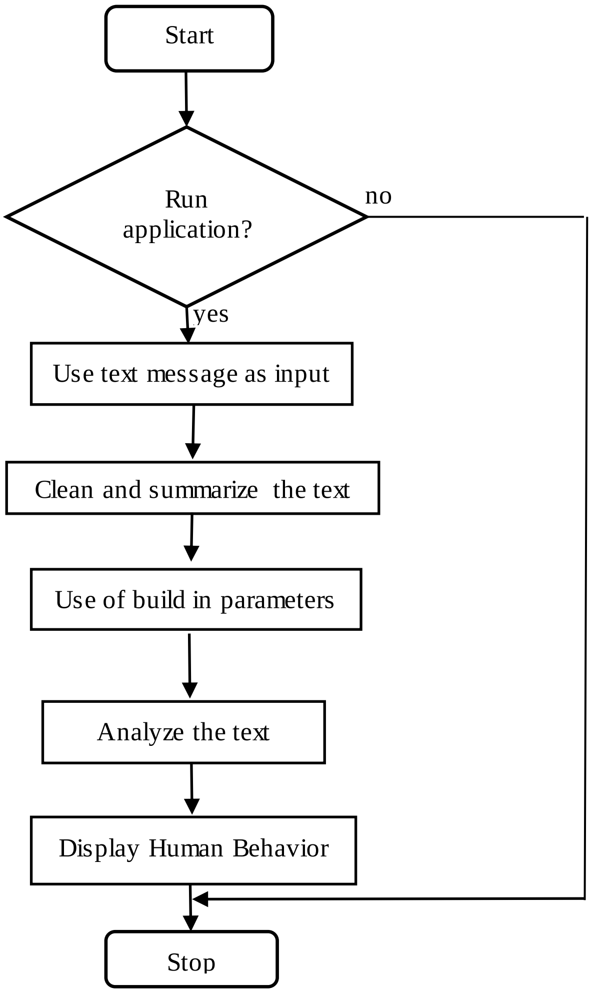
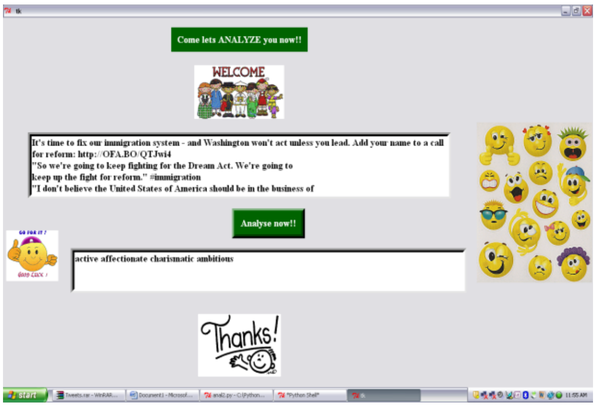

Start
Run application?
no
yes
Use text message as input
Clean and summarize the text
Use of build in parameters
Analyze the text
Display Human Behavior
Stop

Fig. 3: Flow Diagram of this application

IV. Experiment and Analysis

The fig 4 shows the text extracted from Twitter of the most famous powerful politician of US. The text extracted is preprocessed and summarized, and is analyzed using this proposed framework. The result displayed on the user interface tells that the person is active, affectionate, charismatic and ambitious.

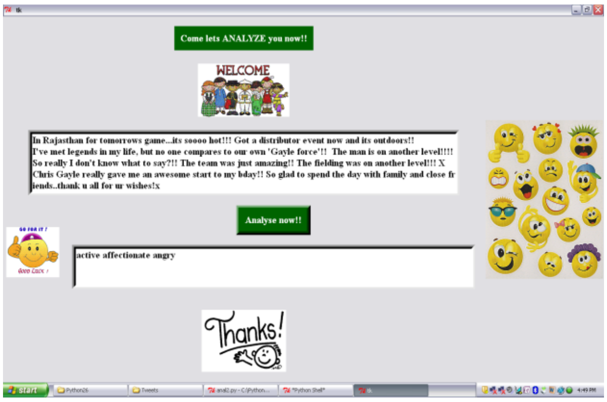


It's time to fix our immigration system - and Washington won't act unless you lead. Add your name to a call for reform: http://OFA.BO/QTJwi4
"So we're going to keep fighting for the Dream Act. We're going to keep up the fight for reform." #immigration
"I don't believe the United States of America should be in the business of

active affectionate charismatic ambitious

Fig. 4: Tweets of a great powerful politician of US

Similarly, the two hundred tweets of a well-known Indian Entrepreneur's son are extracted to determine his personality traits. The outcome of the analysis of his tweets predicted him as active, affectionate and angry.

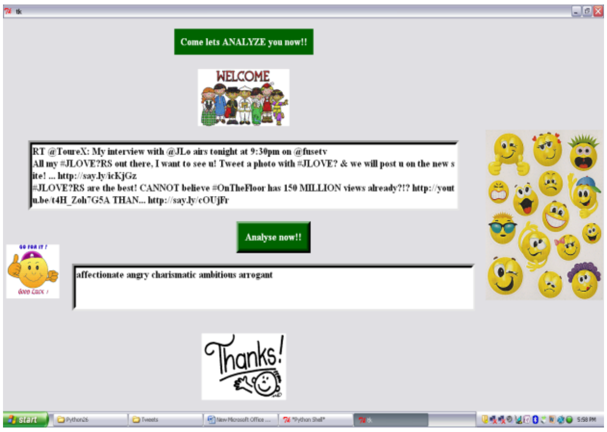


In Rajasthan for tomorrows game...its soooo hot!!! Got a distributor event now and its outdoors!!
I've met legends in my life, but no one compares to our own 'Gayle force'!! The man is on another level!!!!
So really I don't know what to say?!! The team was just amazing!! The fielding was on another level!!! X
Chris Gayle really gave me an awesome start to my bday!! So glad to spend the day with family and close fr
iends..thank u all for ur wishes!x

active affectionate angry

Fig. 5: Tweets of a renowned South Indian Industrialist's Son

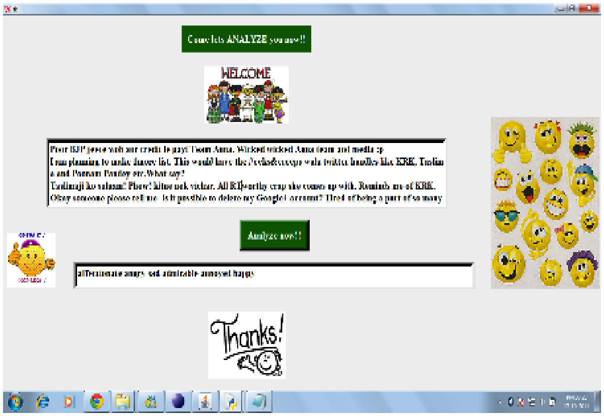
The tweet extracted from a celebrated Hollywood actress when given as input to application, her personality traits are predicted as affectionate, charismatic and ambitious.



RT @ToureX: My interview with @JLo airs tonight at 9:30pm on @fusetv
All my #JLOVE?RS out there, I want to see u! Tweet a photo with #JLOVE? & we will post u on the new s
ite! ... http://say.ly/icKjGz
#JLOVE?RS are the best! CANNOT believe #OnTheFloor has 150 MILLION views already?!? http://yout
u.be/t4H_Zoh7G5A THAN... http://say.ly/cOUjFr

affectionate angry charismatic ambitious arrogant

Fig. 6: Analysis of a famous Hollywood Actress



Poor BJP jeeve wali aur credit le gayi Team Anna. Wicked wicked Anna team and media :p
I am planning to make losers list. This would have the ?echo&creeps wala twitter handles like KRK, Taslim
a and Poonam Pandey etc. What say?
Taslimaji ko salaam! Phew! kitne nek vichar. All KJ-worthy crap she comes up with. Reminds me of KRK.
Okay someone please tell me. Is it possible to delete my Google+ account? Tired of being a part of so many

affectionate angry sad admirable annoyed happy

Priyanka: A common person.

Fig. 7: Tweets of a common close colleague

Also, the tweets of a common known person to authors have also been analyzed and the prediction about her personality is affectionate, annoyed, sad, angry and admirable. This application has been tested on around 50 close known persons also and shown remarkable similarity to their traits.

Also, the tweets of world famous singer Shakira were also extracted and she is predicted as affectionate, charismatic and sad.

The above-predicted results of different people from different culture and background are actually known as the same way in real world also. However, for analyzing human behavior only two hundred tweets are considered. But, more accurate results are expected if these tweets are extracted over a period of time. As on a particular day, a person can be in a bad mood although he is a charming personality.

V. Conclusion and Future Enhancements

This application is aimed at analyzing Human Behavior based on the usage of text. Text plays a fundamental and vital role in determining behavior patterns of Human. As a result, text was extracted from social networking site where most of the common and famous personalities impart day-to-day activities and they exchange words, here the networking site considered is Twitter. This extracted text is an input to the very important phase of the application called text preprocessing. Preprocessed text was conceded so that most relevant and important data are considered. The text was preprocessed and summarized by the following approach.

1. Removal of stop words.
2. Using Hierarchical clustering that is clustering the words if these words in the text are semantically similar else a single word would remain as a single cluster. The semantically similar words are determined using knowledge pool Wordnet [4, 6], which provides synsets and synonyms of words in the text.

The preprocessed text so obtained is the input to the next and final phase called Text Analysis to extract the behavior pattern of allied Human. Text is analyzed based on the psychological parameters that were base line to extract behavior of Human being. These parameters are created as a dictionary. The dictionary has few listed words that would most commonly be classified into it. Wordnet will provide the synonyms and also the synsets of the listed words of the dictionary created. Thus analyze the behavior. The application was applied on the text extracted on famous and great personalities in the real world. The results are accurate and can be mapped to the individual. This application finds its scope in various fields- Political Science, Criminal Science and so on.

In future the following is in progress with respect to the existing framework.

The language that is extracted at present is in English, we can look ahead in considering the rest of the languages. The text is summarized looking into semantically similar words in the text, thus the semantic relatedness of the extracted words is other major scope of this project. Text can be extracted from popular and existing networking sites, Gmail, Yahoo, and other mail accounts, so as to extract mails and know the person-to-person communication and analyze data (text). The dictionary created is static, will be dynamic in future. The concept, context and moods of the allied person would differ the way in which he/she communicates and texts, considering them to analyze the text would aid the result positive and practical. Also, the analysis of tweets for a person can be done over a period of time for more accurate results. It is highly possible that while writing tweets, a person is in some particular state of mind. He can be very sad while being on social networking site for some specific reason but otherwise he may be a very charming and happy personality. So, this direction will also be explored by considering the extraction of tweets over a period of days. Currently, only two hundred tweets are extracted for a person. Still so far the result of the application for different persons is highly promising.

Acknowledgement

We would like to express our gratitude to Dr. J Suryaprasad, Principal and Director, PESIT Bangalore South Campus for his valuable suggestions for preparing Experiment and Analysis Section.

References

[1] Thomas W.M "Data Mining and Text Mining-A Business Applications Approach", Pearson Education, 2008, pp103-123.

[2] Pang-Ning T, Michael S, Vipin K, " Introduction to Data Mining" Published by Addison Wesley, ed 2005

[3] http://en.wikipedia.org/wiki/Natural_Language_Processing

[4] http://en.wikipedia.org/wiki/WordNet

[5] http://computer.howstuffworks.com/web-302.htm

[6] http://wordnet.princeton.edu/wordnet/

[7] George A. Miller "WordNet: A Lexical Database for English". Communications of the ACM vol.38, no.11, 1995, pp 39-41.

[8] http://comminfo.rutgers.edu/~msharp/text_mining.html

[9] Sonia H, Jairo A. Laura M, Andrian M " On the use of automated text summarization techniques for summarizing source code", In proc. Of 17th Working Conf. on Reverse Engineering, 2010, pp35-44.

[10] Jose S, Thomas M " Understanding and Predicting Human Behavior for Social Communities", In handbook of Springerlink Social Network Technologies and Applications, 2010, part 3, pp 427-445.

[11] Jayfun K, Barla C, Cevdet A, Fazli C " Chat Mining for Gender Prediction" In Springer-Verlag LNCS 4243, pp 274-283, 2006.

[12] Jayfun K, Barla C, Cevdet A, Fazli C "Chat Mining: Predicting user and message attributes in computer-mediated communication" In J. of Information Processing and Management, Vol 44, issue 4, july 2008, pp 1448-1466.

[13] Hariharan S, Aashika Rani K.R "Gender Prediction in chat based medium's using text mining" In Intl. j. of Research and reviews in Information Sciences, Vol.1, no. 1, March 2011, pp 18-22.

[14] Mats D " Automatic Prediction of gender political affiliation and age in Swedish politicians from the wordings of their speeches- A comparative study of classifiability", In Oxford J. of Literary and Linguistic Computing, Vol 27, issue 2, pp: 139-153.

[15] Shitala P, Vivek S, Akshay S " Handwritining Analysis based on Segmentation Method for Prediction of Human Personality using Support Vector Machine" In Intl. J. of Computer Applications, Vol 8, no. 12, Oct 2012.

[16] Robert S. Feldman "Understanding Psychology" Published by McGraw-Hill, 6/e.

Ramya Sharada is currently working as Associate IT Consultant with ITC Infotech, Bangalore. She has completed her Bachelor's in Engineering from PESIT, Bangalore South Campus in Information Science Engineering. Her areas of interest are Text processing and Mining, Spatial Data Mining and knowledge based systems.

Arti Arya has completed BSc(Mathematics Hons) in 1994 and MSc(Mathematics) in 1996 from Delhi University. She has completed her Doctorate of Philosophy in Computer Science Engineering from Faculty of Technology and Engineering from Maharishi Dayanand University, Rohtak, Haryana in 2008. Her areas of interest include spatial data mining, knowledge based systems, text mining, unstructured data management, applied numerical methods and biostatistics. She is a life member of CSI and member

IEEE. She is on the reviewer board of many reputed International Journals.

She is currently serving as Professor and Head of the Department (MCA), PES Institute of Technology, Bangalore South Campus. She has more than twelve years of teaching and six years of research experience.

Ragini S is currently working as Associate IT Consultant with ITC Infotech, Bangalore. She has completed her Bachelor's in Engineering from PESIT, Bangalore South Campus in Information Science Engineering. Her areas of interest include Natural Language processing, programming languages, knowledge based systems.

Harish Kumar V is pursuing his post graduate program(MCA) in PES Institute Of Technology(south campus), Bangalore, India. His research areas include text data mining and natural language processing.

Abhinaya G is currently working as Project Engineer with Wipro Technologies, Bangalore. She has completed her Bachelor's in Engineering from PESIT, Bangalore South Campus in Information Science Engineering. Her areas of interest include Natural Language processing, big data handling, database management system.

How to cite this paper: Ramya Sharada K, Arti Arya, Ragini S, Harish Kumar, Abinaya G,"A Text Analysis Based Seamless Framework for Predicting Human Personality Traits from Social Networking Sites", International Journal of Information Technology and Computer Science(IJITCS), vol.4, no.10, pp.37-43, 2012. DOI: 10.5815/ijitcs.2012.10.04