International Journal of Soft Computing and Engineering (IJSCE)
ISSN: 2231-2307, Volume-2, Issue-4, September 2012

ADATPM: Analysis and Design of Author's Trait Processing Module for textual Data

Gaurav Kansal, Maneesha

Abstract- Trait theory is a major approach to the study of human personality. Personality is the branch of psychology which is concerned with providing a systematic account of the ways by which we can differentiate one-another. Individuals differ from one another in a variety of ways: their anatomical and physiognomic characteristics, their personal appearance, grooming, manner of dress, their social backgrounds, roles and other demographic characteristics, their effect on others or social stimulus value and their temporary states, moods, attitudes and activities at any given moment in time.. In this paper we have designed a system that takes text input and returns the author's trait accordingly. Since human tendencies are largely dependent on environmental and situational consistencies, we have considered five different traits in our identification. These are High Extrovert, Low Extrovert, High Introvert, Low Introvert and Ambivert. Our algorithm refines the author's text under eight different properties. The text undergoes POS tagger where each word is assigned a tag. After analyzing the tag we generate Feature Vector matrix (FVM), we use this FVM for our analysis as well as for the classification. We have applied our proposed algorithm on different 280 files. These files are also annotated by human. We compare the result got from human annotation and proposed algorithm and we found that the accuracy of our algorithm is 84.26%.

Index Terms—FVM, POS, SVM, Trait Theory.

I. INTRODUCTION

In this world there are about 70,000 million peoples and every people is different with each other with respect to some individual features. The set of these features are known as personality. Personality is the area of psychology which mainly concerned with providing a systematic account of the ways in which every person differs from one and another. Individual differs under varieties of ways physiognomic characteristics; grooming, manner of dress, their social backgrounds, roles, and other demographic characteristics; their effect on others or social stimulus value and their anatomical. There are some other features which differ individual's e.g. at any given moment in time, their temporary states, moods, attitudes and activities. The study of personality always provide a systematic account of individual differences in human tendencies to act or not to act in certain ways on certain occasions, these tendencies are proclivities, propensities and dispositions, inclinations. Generally these tendencies are also known as Traits. A trait is what we call a characteristic way in which an individual perceives feels, believes, or acts. When we casually describe someone, we are likely to use trait terms. Psychologists, especially personologists, are very interested in traits. They are especially interested in finding which traits are broad and possibly genetically based, as opposed to ones that are rather peculiar and can change easily.

Over the years, we have had a number of theories that attempt to describe the key traits of human beings. There are so many factors for indentifying a person e.g. name, place, trait as well as personality. There are various attributes of a person – behavior, temperamental, emotional and mental. These attributes characterize a unique individual. Humans have the propensity to explain the other humans' behavior in terms of even properties that are variously mixed on the basis of observation of everybody behavior. Today's time is the time of information and internet. The best way for sharing the information is e-mail, blog, online diaries etc. The text written by a person is also reflect the personality of author's and emotion of the author's as well as after analyzing the text we can also conclude that what a person written that is positive or negative and what is the intensity of his/her statement/text.

In Natural Language Processing we generally processed the natural language; here the meaning of natural language is any common language in which peoples are sharing the thought. There is an area of Natural Language Processing known as Emotion Mining, under this area we can find the emotion and traits of the author's as well as emotion present in the text written by the Author's. Finding of the emotion and trait of author's is a very challenging task because thought of a person extensively dependent on the atmosphere as well as scenario of that time when did he write the text.

In psychology, Trait theory is a major approach to the study of human personality. Trait theorists are primarily interested in the measurement of traits, which can be defined as habitual patterns of behavior, thought, and emotion. According to this perspective, traits are relatively stable over time, differ among individuals (e.g. some people are outgoing whereas others are shy), and influence behavior.

Gordon Allport was an early pioneer in the study of traits, which he sometimes referred to as dispositions. In his 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 emphases "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. Hans 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 are sufficient. Virtually all trait models, and even ancient Greek philosophy, include extraversion vs.Introversion as a central dimension of human personality. Another prominent trait that is found in Nearly all models are Neuroticism, or emotional instability.

Manuscript received on September, 2012
Gaurav Kansal, E&R, Infosys.
Maneesha, E&R, Infosys

Eysenck was one of the first psychologists to Study personality with the method of factor analysis, a statistical technique introduced by Charles Spearman. Eysenck's results suggested two main personality factors. The first factor was the tendency to experience negative emotions, and Eysenck referred to it as Low Extrovert. The second factor was the tendency to enjoy positive events, especially social events, and Eysenck named it High Extrovert. Similarly High Introvert always use negative emotion adjective with their property e.g. use of short sentences. Low Introvert always try to use positive adjective but some additional property e.g. use of articles etc. The two personality dimensions were described in his 1947 book Dimensions of Personality. It is common practice in personality psychology to refer to the dimensions by the first letters, E and N. E and N provided a 2-dimensional space to describe individual differences in behavior. An analogy can be made to how latitude and longitude describe a point on the face of the earth. Also, Eysenck noted how these two dimensions were similar to the four personality types first proposed by the Greek physician Hippocrates.

The major strength of Eysenck's model was to provide detailed theory of the causes of Personality. For example, Eysenck proposed that extraversion was caused by variability in cortical arousal: "introverts are characterized by higher levels of activity than extraverts and so are chronically more cortically aroused than extraverts". While it seems counterintuitive to suppose that introverts are more aroused than extraverts, the putative effect this has on behavior Is such that the introvert seeks lower levels of stimulation. Conversely, the extravert seeks to heighten his or her arousal to a more optimal level (as predicted by the Yerkes-Dodson Law) by increased activity, social engagement and other stimulation-seeking behaviors.

One of the long held goals of psychology has been to establish a Model that can conveniently describe human personality and disorders therein, with the intent to use this model in the remedying of personality disorders and improving general understanding of personality. Currently, a handful of models have risen to prominence, and have thus far stood the test of time. Some models are more generally accepted than others. Support for some models seems to come and go in cycles.

II. RELATED WORK

We can identify the trait of person after analyzing gesture, voice communication as well as with the help of written text. The use of term "Trait" in contemporary psychological discourse carries with it implications of a particular theoretical commitment, a preferred method of scientific investigation, and a philosophical preference for certain kinds of explanation in theory construction. Hence, it is necessary to make it clear at the outset that an interest in human tendencies (traits) does not imply a theoretical pre commitment to such issues as whether traits are manifestations of generative or causal mechanisms. The identification of Author's trait is very important and useful for various purposes e.g. in Medical, mental status etc.

In the research paper named as "**Allport's Theory of Traits--A Critical Review of the Theory and Two Studies**" written by Louise Barkhuus, Patricia Csank, here author reviews Gordon Allport's theory of traits as well as two of his studies, "Personality Traits", 1921 and "Letters from Jenny", 1966. His theory, which is based more on his view of human nature than on research, distinguishes between common traits and individual traits, with emphasis on the Individual traits. The two studies illustrate how Allport applies the theory in his research. Finally the paper concludes that although Allport's trait theory only capture parts of the concept of personality, credit should be given due to the fact that the theory is an early attempt to describe and measure personality. Gordon W. Allport (1897–1967) was the first psychologists who gave thorough thought to the concepts of traits. He developed his own trait theory and he continued to view the trait as the most appropriate way of describing and studying personality. He is, by many, actually considered to be the first psychologist dealing with personality at all and was the first to offer a class in this field at Harvard University in 1924 (Schultz, 1976; Pervin& John, 1997). Throughout his life, Allport continued to develop and work with his trait theory and he inspired many other psychologists who also adopted this approach to personality or developed their own trait theory (e.g. Eysenck, McClelland). The aim of this paper is to review Allport's trait theory as described in his own published material supplemented by comments from other scholars. The paper's focus is on the theory of traits and Allport's view of personality. Although much literature has been published on the concept of personality traits, seen from other perspectives, this will not be dealt with. Allport's other aspects of personality psychology will only be mentioned briefly or in connection to his trait theory.

In order to understand Allport's theory of traits, it is important to know how he approached Psychology and in particular the issue of personality. In many ways, his views were opposite from the ones of the psychoanalysts but they were also very different from the behaviorists. Allport viewed psychology as the study of the healthy person. He believed, in contrast to for Example the psychoanalysts, that studying the healthy personality is much different and incompatible with that of the pathological personality (Schultz, 1976). Another basic approach he takes is that of the individual human as unique. Each person is different from the other and should therefore be studied accordingly. Individuals can still be compared but Allport's understanding of psychology goes beyond just comparison. He emphasizes this individuality in virtually all aspects of his psychology, another contrast to the view of the psychoanalysts as well as other psychologists, who put emphasis on similarities within people (Chaplin &Krawiec, 1968). Another radical view of Allport is one regarding the dynamics within the individual. He referred to this as functional autonomy. This aspect of his psychology is probably where Allport differs most from other psychologists of his time, especially psycho-analysts like Freud and Jung but also behaviorists like Skinner (Chaplin &Krawiec, 1968). Allport believes that motivation occurs independent of past experiences. It is the present motives such as interests, attitudes and life style that govern a person's behavior. He stresses the close relationship between motives and cognitive Processes and argues that all motives are a combination of these. This way the individual's "cognitive style" is affected by the individual's self-perception and only indirectly affected by his/her past. We shall later see how the trait theory relates to this concept of motivational autonomy. Keeping these basic approaches in mind, Allport's theory of traits seems a natural part of his description of personality. We shall now see how he explained traits as the core of personality. Allport defines a trait as "a generalized and focalized neuropsychic system

(peculiar to the individual), with the capacity to render many stimuli functionally equivalent, and to initiate and guide consistent (equivalent) forms of adaptive and expressive behaviour" (Allport, 1937, p.295).First one notices that Allport describes a trait as a neuropsychic system. He firmly believes that traits are real and exist within the person. Allport does not mean that a trait is what we today would call genetic, although he does regard some traits as "hereditary" (Pervin& John, 1997). He means that the traits make behavior consistent and that a trait is still there even if there is no one around to see it. In his book "Personality – A psychological interpretation" from 1937, Allport uses the example of Robinson Crusoe and asks the provocative question: "Did Robinson Crusoe lack traits before the advent of Friday?" (Allport, 1937, p. 289). Still traits can be evoked by a certain social situation. This issue will also be dealt with when discussing the inter-dependence of traits.

In the research paper named as **A "Big Five" Scoring System for the Myers-Briggs Type Indicator** written by Robert J. Harvey, William D. Murry, Steven E. Markham discussed about the degree to which personality tests have been used as employee selection and placement tools has varied considerably. After enjoying a period of popularity during the earlier part of this century, during the 1960's the prevailing view (e.g., Guion&Gottier, 1965) shifted to a much more negative assessment: namely, that "the validity of standard personality measures for personnel selection was so poor that their continued use seemed unwarranted" (Hogan, 1991, p. 896). However, in more recent years personality-based employee selection tests have staged a Resurgence in popularity, spurred by the appearance of empirical studies and meta-analyses that supported their utility as assessment devices (e.g., Barrick& Mount, 1991; Hough, Eaton, Dunnette, Kamp, &McCloy, 1990; Mount, Barrick, & Strauss, 1994; Schmit& Ryan, 1993; Tett, Jackson, & Rothstein, 1991). One factor that has energized and directed research and practice in this area has been the growing acceptance of the Big Five view of the structure of personality (e.g., Cortina, Doherty, Schmitt, Kaufman, & Smith, 1992; Digman, 1990; Hogan & Hogan, 1992; McCrae & Costa, 1987; Schmit& Ryan, 1993). According to the Big Five taxonomy, the primary dimensions of personality are Extraversion, Agreeableness, Conscientiousness, Emotional Stability, and Openness to Experience; Although debate continues regarding the question of which Big Five scales are the most generally useful in selection contexts (e.g., Barrick& Mount, 1991; Tett et al., 1994; Ones, Mount, Barrick, & Hunter, 1994; Tett, Jackson, Rothstein, &Reddon, 1994) -- as well as thequestion of whether subscales of the Big Five provide higher levels of predictability than the main scales (e.g., Hogan & Hogan, 1992; Hough et al., 1990) -- it is evident that the Big Five taxonomy has exerted a major positive impact on current uses of personality tests for employee selection. Despite the fact that it was not developed in the Big Five tradition, the Myers-Briggs Type Indicator (MBTI; Briggs & Myers, 1976; Myers & McCauley, 1985) has enjoyed widespread popularity in applied organizational contexts. Indeed, by some estimates the MBTI has become the most widely used personality assessment instrument in corporate America, with an estimated 1.5 million workers having completed the MBTI in 1986 alone (Moore, 1987); in 1991, that estimate had risen to over 2 million people (Suplee, 1991). The MBTI is used in a wide variety of organizational applications: for example, Poilitt (1982) described the use of the MBTI for career guidance and personal development;

Hartzler and Hartzler (1982) described the application of the MBTI for "planning, organizing, directing, and controlling" (p. 20) the actions of other workers; Garden (1989) used MBTI profiles to predict employee turnover; Gauld and Sink (1985) and Sample and Hoffman (1986) described the use of the MBTI for organizational development; and several studies (e.g., Gough, 1976; Hall & MacKinnon, 1969; Kirton, 1976) have used the MBTI to predict aspects of job performance (in these examples, creativity and innovation). The MBTI has even found application in job analysis and synthetic test validation: based on a job's Position Analysis Questionnaire (PAQ; McCormick, Jeanneret, &Mecham, 1972) profile, an estimate of the MBTI profile one would expect to find among job incumbents can be produced by the PAQ's scoring service using synthetic validity (e.g., Jeanneret, 1992; Mecham, 1989).

In the research paper named as **"What Are They Blogging About? Personality, Topic and Motivation in Blogs"** written by Alastair J. Gill. Here authors discussed about the personality of author on blog data. Blog is a place in Internet where a person shares his views about any entity. Personal weblogs or we can say it blogs provide the individuals with the opportunity to write freely and express themselves online in the presence of others. In this paper authors examine the content of blogs to provide the insight into the role of personality in motivation for blogging. As predicted, we find that highly Neurotic authors use blogs to serve a cathartic or auto-therapeutic function, and reflect mainly upon themselves and negative emotions. Highly Extraverted blog authors, as expected, use blogs to document their lives at a high level, and uniquely interact directly with the reader. Additionally Extraverts use blogs to vent both positive and negative emotions. Bloggers who are high scorers on the Openness trait are more concerned with leisure activities, although they are more evaluative than intellectual, whereas high Conscientiousness bloggers tend to report daily life – and work – around them. As in other contexts, expressing positive rather than negative emotions is associated with high Agreeableness, but that trait is associated with self reference to a greater degree in blogs than elsewhere. In general, findings are consistent with other contexts indicating that bloggers tend to adapt to the possibilities of the medium, rather than try to present themselves differently. In this paper author also discuss about the properties of different personality traits which as follows:

(A). Neuroticism: Blogs authored by high Neurotics are more likely to serve a cathartic or auto-therapeutic purpose. This is likely to overlap with findings from previous literature, namely, (a) greater self reference (first person singular pronoun) and negative emotion words (Pennebaker and King, 1999), and (b) fewer references to others (second, or third person pronouns) (Oberlander and Gill, 2006). Additionally from previous blog findings, we expect (c) the topic to focus more on jobs and physical states (Nowson, 2006).

(B). Extraversion: We expect high Extravert blog authors to write blogs more concerned with documenting life, with this characterised by (a) more verbs (past, present and future), and time references. Consistent with previous literature, it is likely that Extraverts will use (b) more pronouns (first, second and third person) (Pennebaker and

King, 1999). Additionally (c) we expect fewer negative emotion words (Pennebaker and King, 1999).

(c). Openness: High Openness bloggers are likely to write blogs reflecting their interest, opinions or feelings. We therefore expect (a) topics to focus on leisure activities, and (b) a greater number of cognitive mechanism words and words concerned with the senses. From previous literature, we expect (c) fewer first person singular pronouns and present tense verbs (Pennebaker and King, 1999), and fewer references to occupation and more positive emotion words (Nowson, 2006).

(D). Conscientiousness: We expect highly conscientious bloggers to write about their interests and to (a) use more words relating to their occupation, and also to time, and past, present and future verbs. We also expect them to (b) use more positive emotion words and fewer negative emotion words (Pennebaker and King, 1999).

(E). Agreeableness: We predict that this trait will mainly influence what topics the author chooses to write about or avoid in their blog. Following previous literature, we expect (a) fewer negative and more positive emotion words, and more self references (Pennebaker and King, 1999). We also expect (b) fewer bodily references (Nowson, 2006).

In the paper named as **"More Blogging Features for Author Identification"** written by HaythamMohtasseb and Amr Ahmed, here authors presented a novel implementation in the field of authorship identification in personal blogs. The improvement is done by utilizing the hybrid collection of linguistic features that best capture the style of users in dairies blogs. Here authors used the features set contain LIWC with its psychology background a collection of syntactic features & part of speech (POS) and the misspelling errors features.

Furthermore, authors analyzed the contribution of each feature set on the final result and compare the outcome of using different combination from the selected feature sets. Here authors create a new category of misspelling words which are mapped into numerical features, are noticeably enhancing the classification results. The paper also confirms the best ranges of several parameters that affect the final result of authorship identification such as the author numbers, words number in each post, and the number of documents/posts for each author/user. The results and evaluation show that the utilized features are compact, while their performance is highly comparable with other much larger feature sets.

In this paper, authors presented research of identifying the bloggers in online diaries by mining their diaries text. We identify the nature and properties of the textual content used by bloggers and find out the superlative collections of linguistic features that best capture the style of authors. In this framework, a large spectrum of experiments have been executed, exploring the significant parameters ranges of the users' number, posts sizes and lengths, and indicating the best set of features that improve the identification percentage. While previous studies in authorship identification achieved high classification accuracy but in different corpus types, we also acquire, according to specific criteria, superior results using a smaller number of features (129) compared to their features numbers. Here authors found that LIWC is the best individual option among other feature sets as a baseline selection. This is due to its dictionary richness which covers a large variety of real life topics that is highly correlated with the content of the diaries blogging text. In additions to the other features sets, the syntactic & POS, which are also improving the result, our created set of misspelling features is enhancing the final outcome of the authorship identification framework. Although previous studies utilized misspelling features, but we chose a very small number of features than their features size, considered the common misspelling errors happened in the diaries, and effectively introduced a new categorization map between the features and the misspelling words.

In the research paper named as **"Whose thumb is it anyway? Classifying author personality from weblog text"** written by Jon Oberlander and Scott Nowson report on initial results on the relatively novel task of automatic classification of author personality. Using a corpus of personal weblogs, or 'blogs', they investigate the accuracy that can be achieved when classifying authors on four important personality traits. We explore binary and multiple classifications, using differing sets of n-gram features. Results are promising for all four traits examined.

In this paper they have discussed about Cattell's pioneering work led to the isolation of 16 primary personality factors and later work on secondary factors led to Costa and McCrae's five factor model, closely related to the 'Big Five' models merging from lexical research (Costa and McCrae, 1992). Each factor gives a continuous dimension for personality scoring. These are: Extraversion; Neuroticism; Openness; Agreeableness; and Conscientiousness (Matthews et al., 2003). Work has also investigated whether scores on these dimensions correlate with language use (Scherer, 1979; Dewaele and Furnham, 1999). Building on the earlier work of Gottschalk and Gleser, Pennebaker and colleagues secured significant results using the Linguistic Inquiry andWord Count Text analysis program (Pennebaker et al., 2001). This primarily counts relative frequencies of word-stems in pre-defined semantic and syntactic categories. It shows, for instance, that high Neuroticism scorers use: more first person singular and negative emotion words; and fewer articles and positive emotion words (Pennebaker and King, 1999). So, can a text classifier trained on such features predict the author personality? We know of only one published study: Argamon et al. (2005) focused on Extraversion and Neuroticism, dividing Pennebaker and King's (1999) population into the top- and bottom-third scorers on a dimension, and discarding the middle third. For both dimensions, using a restricted feature set, they report binary classification accuracy of around 58%: an 8% absolute improvement over their baseline. Although mood is more malleable, work on it is also relevant (Mishne, 2005). Using a more typical feature set (including n-grams of words and parts-of-speech), the best mood classification accuracy was 66%, for 'confused'. At a coarse grain, moods could be classified with accuracies of 57% (active vs. passive), and 60% (positive vs. negative). So, Argamon et al. used a restricted feature set for binary classification on two dimensions: Extraversion and Neuroticism. Given this, we now pursue three questions. (1) Can we improve performance on a similar binary classification task? (2) How accurate can classification is on the other dimensions? (3) How accurate can multiple— three-way or five-way—classification be? In this paper authors used Support Vector Machine for the binary sentiment classification task. This paper has reported the first stages of their investigations into classification of author personality from weblog text. Results are quite promising and comparable across all four personality traits. It seems that even a small selection of

features found to exhibit an empirical relationship with personality traits can be used to generate reasonably accurate classification results. Naturally, there are still many paths to explore. Simple regression analyses are reported in Nowson (2006); however, for classification, a more thorough comparison of different machine learning methodologies is required. A richer set of features besides n-grams should be checked, and we should not ignore the potential effectiveness of unigrams in this task (Pang et al., 2002). A completely new test set can be gathered, so as to further guard against over fitting, and to explore systematically the effects of the amount of training data available for each author. And as just discussed, comparison with human personality classification accuracy is potentially very interesting. However, it does seem that we are making progress towards being able to deal with a realistic task: if they spot a thumbs-up review in a weblog, they should be able to check other text in that weblog, and tell whose thumb it is; or more accurately, what kind of person's thumb it is, anyway. And that in turn should help tell us how high the thumb is really being held.

In the research paper named as **"Support Vector Machines Classification with a Very Large-scale Taxonomy"** written by Tie-Yan Liu, Yiming Yang, Hao Wan, Hua-Jun Zeng, Zheng Chen, and Wei-Ying Ma, here authors discussed about the classification and regression of the entity with the help of Support Vector Machine. Here they presented that Very large-scale classification taxonomies typically have hundreds of thousands of categories, deep hierarchies, and skewed category distribution over documents. However, it is still an open question whether the state-of-the-art technologies in automated text categorization can scale to (and perform well on) such large taxonomies. In this paper, they report the first evaluation of Support Vector Machines (SVMs) in web-page classification over the full taxonomy of the Yahoo! categories. Our accomplishments include: 1) a data analysis on the Yahoo! taxonomy; 2) the development of a scalable system for large-scale text categorization; 3) theoretical analysis and experimental evaluation of SVMs in hierarchical and non-hierarchical settings for classification; 4) an investigation of threshold tuning algorithms with respect to time complexity and their effect on the classification accuracy of SVMs. they found that, in terms of scalability, the hierarchical use of SVMs is efficient enough for very large-scale classification; however, in terms of Effectiveness, the performance of SVMs over the Yahoo! Directory is still far from satisfactory, which indicates that more substantial investigation is needed.

According to their categorization of SVMs in the previous section, for flat SVMs, each SVM model is trained to distinguish one category from all the other categories. For the testing phase, an exhaustive search is used to classify an instance into the category with the highest confidence score. It is clear that the complexity of flat SVMs is proportional to the number of categories. Therefore, when handling hundreds of thousands of categories, the computational load will increase to unacceptable levels. To tackle this problem, people have utilized the hierarchical structure of the taxonomy tree to decompose the classification task. In [32] and [33], Dumais used hierarchical SVMs to classify the Look Smart dataset. For the training phase, a classifier was trained to distinguish only those categories with the same parent node in the taxonomy tree. And for testing, a pachinko-machine search was used, where an SVM model is used only if the model of its parent category says YES on the test instance. They claimed improved classification performance with a significant (i.e. more than 80%) reduction in computation compared to the flat baseline. However, because they only used the top two levels of the LookSmart categories (163 categories in total) in their experiments, their conclusions might not easily generalize to the case of classifying hundreds of thousands of categories. Their previous work, [33], is the first paper to give a theoretical analysis of the scalability of TC algorithms. Using the power law to model the category distributions, they derived the bounds of complexity for both flat and hierarchical SVMs. Experiments were conducted on OHSUMED [34] to verify the theoretical analysis: for example, it took 102 hours to train flat SVMs over OHSUMED and only took 26.3 minutes to train hierarchical SVMs. However, these experiments were not conducted over the full domain of OHSUMED (with 14,321 categories in total) but projected from 94 categories in the heart-disease sub domain. Furthermore, the classification performance was not reported, so the trade-off between effectiveness and efficiency was not discussed. Besides the aforementioned work, other work has also been proposed to investigate the problem of SVM classification over hierarchical taxonomies [36][32][33][37][38][39]. Once again, they verified their findings over datasets with only hundreds, or at most a few thousand categories (such as Reuters 21578, RCV1, the heart-disease sub tree of OHSUMED, and WIPO-alpha [43]). So in summary, the question still remains open as to whether SVMs can scale to hundreds of thousands of categories, and what the tradeoff between efficiency and effectiveness will be. In this regard, it will SVM classification over the full domain of a very large-scale data corpus, which is the motivation of our paper. They also show that the difficulties in applying text categorization algorithms to very large problems, especially large-scale Web taxonomies, have been underestimated or at least not studied thoroughly in the literature. In order to gain a better understanding, we conducted the first evaluation of SVMs with the full Yahoo! web-page taxonomy, which yielded the following new conclusions: 1) Threshold tuning (SCut in our paper) dominates the time complexity of offline training of SVMs, which was not well understood until this study. 2) In terms of scalability, while the complexity of flat SVMs is too high, hierarchical SVMs are efficient enough for very large-scale real-world applications. 3) In terms of effectiveness, neither flat nor hierarchical SVMs can fulfill the needs of classification of very large-scale taxonomies. 4) The skewed distribution of the Yahoo! Directory and other large taxonomies with many extremely rare categories makes the classification performance of SVMs unacceptable. More substantial investigation is thus needed to improve SVMs and other statistical methods for very large-scale applications.

In the research paper named as "Identifying more bloggers: Towards large scale personality classification of personal weblogs" written by Scott Nowson, Jon Oberlander, here the authors have discussed about the identification of authors personality on blog data. Here they reported new results on the relatively novel task of automatic classification of blog author personality. Promisingly high classification accuracies have recently been reported for four important personality traits (Extraversion, Neuroticism, Agreeableness and Conscientiousness). But the blog corpus used in that work

required careful preparation, and was consequently quite small (with less than a hundred authors; and less than half a million words). Here, they provide an initial report on the classification accuracies that can be achieved when classifiers conditioned on the small corpus are applied to a larger, automatically-acquired blog corpus, using lower granularity personality data and substantially less manual preparation (with over a thousand bloggers, and approximately five million words). Predictably, results on the larger corpus are not as impressive as those on the smaller; nevertheless, they point the way forward for further work. In this paper they show that noise in the text give hopeless results, so more automatic processing required handling the larger corpus.

In the research paper named as **"Improving gender classification of blog authors"** written by Argon Mukharjee and Bing Liu here authors discussed the problem of automatically classifying the gender of a blog author has important applications in many commercial domains. Existing systems mainly use features such as words, word classes, and POS (part-of speech) n-grams, for classification learning. In this paper, authors propose two new techniques to improve the current result. The first technique introduces a new class of features which are variable length POS sequence patterns mined from the training data using a sequence pattern mining algorithm. The second technique is a new feature selection method which is based on an ensemble of several feature selection criteria and approaches. Empirical evaluation using a real-life blog data set shows that these two techniques improve the classification accuracy of the current state-of the-art methods significantly.

III. PROPOSED METHODOLOGY

We will discuss our work under following author's trait.

High Extrovert: All those human which belong to this trait category mainly used more words and these words always referencing to themselves and others, as well as words with positive emotions and express more about the certainty while writing an essay.

These type of personality always show greater complexity including with increased use of introducing clause-initial connectives such as then, which and what conjunctions and adjectives for writing a E-mail.

While writing a blogs such type of person uses more present tense verbs with talk of communication.

Low Extrovert: Person belongs to this category always use more negations and negative emotion adjective. Person belongs under this category use articles with greater tentatively. All the above features with respect to essay writing.

In blog writing Low Extrovert person talks about achievements and use words relating to discrepancies.

High Introvert: Person belongs to this category high scorers in monologue situations have been found to use singular and negative emotion words with more first person pronoun. Beside it also use greater talk about discrepancies, jobs and physical states.

Additionally they use less outward-looking discourse, containing fewer phrases referring to others. These persons used more exclusive, inclusive connectives with a greater use of multiple punctuation expressions in essay writing.

Low Introvert: Low Introvert person always refers more to other people and use more nouns and adverbs. While writing in essays high Openness scorers use more articles, longer words and insight words, and fewer first person singular, present tense.

In Blog writing use more longer words and also express positive feelings. They also use fewer negations and write less about the school.

Ambivert: A person who does not belong to any category and categories are High Extrovert, Low Extrovert, High Introvert and Low Introvert then that person belongs to Ambivert.

3.1 Framework

In this project we are trying to identifying the trait of a person who has written some text, so firstly we must design software which takes the text paragraph as input. This inputted paragraph is analyzed by the proposed algorithm. For checking the accuracy we apply Support Vector Machine.

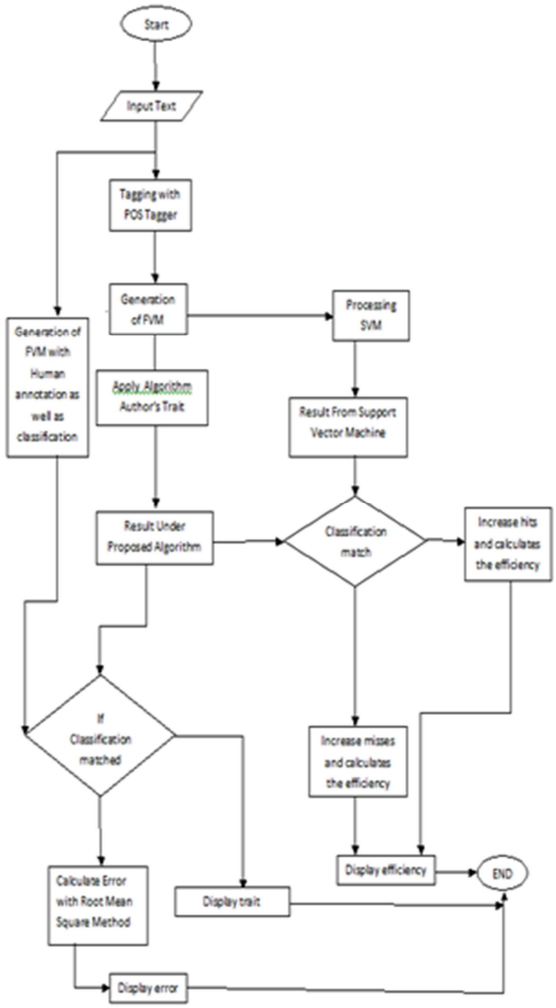


Fig. Architecture of proposed framework

3.2 Proposed Algorithm

We have discussed categorizing the authors on the basis of writing an essay. We have categorized the Author under five categories, after analysis we concluded that following property mainly helps to identify the author's trait.

First person pronoun
Negative emotion adjective
Short sentences
Third person pronoun
Positive emotion adjective
Determiner
Present tense
Past Tense

There are two cases: In Case 1, we can check that either author is High Introvert, Low Extrovert Ambivert. Case 2 we can check that either the author is High Extrovert, Low Introvert Ambivert.

Case 1:

If (60% adjectives have negative emotion)
If (Maximally use first person pronoun)
If (Maximum Short sentences)
Then "highly introvert"
Else "Ambivert"
Else "Low Extrovert"
Else "Ambivert"

Case2:

If (60% adjectives have positive emotion)
If (Maximum present tense)
If (past tense + Present tense >50% of total tenses)
If (more than 60% pronouns are third person pronoun)
Then Author is "High extrovert"
Else "Ambivert"
Else "Ambivert"
Else If (at least 10% of words from articles)
If (60% pronouns from 1st person pronoun)
Then "Low Introvert"
Else "Ambivert"
Else "Ambivert"
Else "Ambivert"

3.3 Tagging the Text

Previously we have proposed algorithm for identifying the trait of author, We can see that there are eight features which are required for categorization. We can obtain these features when we tag the text. The basic role of tagger is to tag all the words of text.

3.4 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 eight 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.

3.5 Classification with Support Vector Machine (SVM)

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. Since an SVM is a classifier, then given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that predicts whether a new example falls into one category or the other. 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>

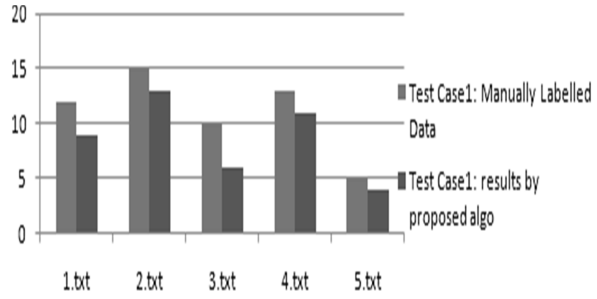
<label> is the target value of the training data. For classification, it should be an integer which identifies a class (multi-class classification is supported). The indices must be in an ascending order. The labels in the testing data file are only used to calculate accuracy. If they are unknown, just fill this column with a number. For giving input in the Support Vector Machine we generate a Feature Vector Matrix of the inputted text. There are basically two functions of Support Vector Machine first one is to train the machine. For executing the training part we must have thecorpus. Property of this corpus is that it belongs to a particular class and it computed previously. After training the machine we pass our input text as test file. After applying the machine on both file it generates the result and it shows that our test file belongs to which category.

In this project we have used binary classification. Binary classification means that test data belongs to a particular class or not, in our project there are 5 classes so we must apply the multiclass classification of Support Vector Machine. We can also apply multiclass feature with the help of binary classification. It can be applied in the way that we can calculate the accuracy with respect to all the classes, for a particular class our result will be the best then we can say that test file belongs to that particular class.

IV. RESULTS

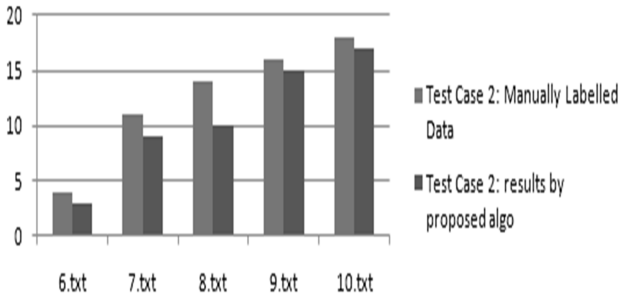
We have executed the algorithm on various files. Here we first manually tag all paragraph of file and it is assumed that manual tagging is 100% correct, after then we have passed these individual paragraph as input in our proposed algorithm. The various Test Cases are as follows:

Test Case 1 as follows:



In Test Case 1 There are five files 1.txt, 2.txt, 3.txt, 4.txt, 5.txt all these files are the collections of paragraph e.g. 1.txt has twelve paragraph. Here we first manually tag the entire twelve paragraphs and we have passed these entire twelve paragraphs as input in our algorithm and we found that nine out of twelve results are matched with manual tagging. Similarly we repeat the above mentioned process for other files e.g. 2.txt, 3.txt etc.

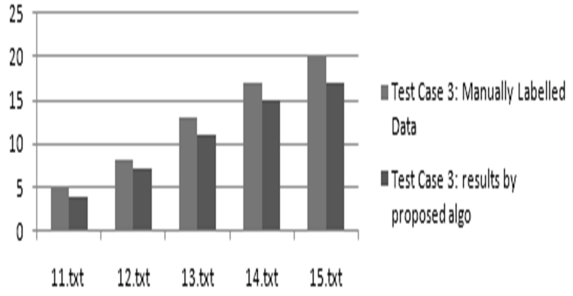
Test Case 2 as follows:



In Test Case 2 There are five files 6.txt, 7.txt, 8.txt, 9.txt, 10.txt all these files are the collections of paragraphs e.g. 9.txt has Sixteen paragraphs. Here we first manually tag the entire sixteen paragraphs and we have passed these entire

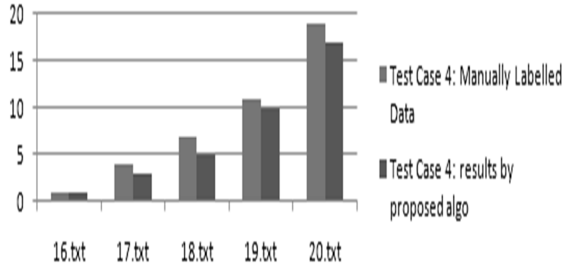
sixteen paragraphs as input in our algorithm and we found that Fifteen out of Sixteen results are matched with manual tagging. Similarly we repeat the above mentioned process for other files e.g. 6.txt, 10.txt etc.

Test Case 3 as follows:



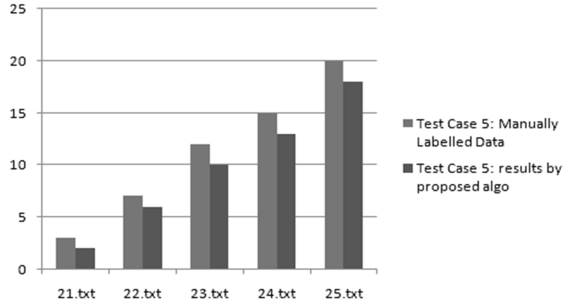
In Test Case 3 There are five files 11.txt, 12.txt, 13.txt, 14.txt, 15.txt all these files are the collections of paragraph e.g. 15.txt has Twenty paragraph. Here we first manually tag the entire sixteen paragraphs and we have passed all these twenty paragraphs as input in our algorithm and we found that Seventeen out of Twenty results are matched with manual tagging. Similarly we repeat the above mentioned process for other files e.g. 14.txt, 15.txt etc.

Test Case 4 as follows:



In Test Case 4 There are five files 16.txt, 17.txt, 18.txt, 19.txt, 20.txt all these files are the collections of paragraph e.g. 18.txt has Seven paragraph. Here we first manually tag all the seven paragraphs and we have passed these entire seven paragraphs as input in our algorithm and we found that Five out of Seven results are matched with manual tagging. Similarly we repeat the above mentioned process for other files e.g. 14.txt, 15.txt etc.

Test Case 5 as follows:



In Test Case 5 There are five files 21.txt, 22.txt, 23.txt, 24.txt,25.txt all these files are the collections of paragraphs e.g. 21.txt has Three files. Here we first manually tag the entire three paragraphs and we have passed all these three files as input in our algorithm and we found that two out of three results are matched with manual tagging. Similarly we repeat the above mentioned process for other files e.g. 22.txt, 23.txt etc.

V. CONCLUSION

In the proposed work we have designed a framework for identifying the author's trait. The writing style of individual category is different, but this different is very minute. In our project we have taken five traits and we categories these five traits on the basis of eight features. We have applied our proposed frame work on 280 different textual data; these textual data are manually tagged. We compare result drawn from algorithm and manual tagged data and we found that for results of 236 files are same, so we conclude that the accuracy of proposed algorithm is 84.28%. In our project it is a great challenge to find the all eight feature of the text. We have applied Part of Speech Tagger for finding the entire feature. The basic role of tagger is to tag all the words exist in the text. After tagging all the words we generate the feature vector matrix. The role of feature vector matrix in our project is very important because categorization will be occurred on this feature vector matrix. Feature vector matrix is tabular representation of all the feature of text and these features will be used for categorization the text. For categorization we have used the Support Vector Machine. This feature vector matrix is passed in Support Vector Machine as input and Support Vector Machine categorize the text. The behavior of author is changed rapidly according to the atmosphere. Here atmosphere mean circumstances around the author e.g. Weather, personal problem etc. So it is very difficult to identify the exact nature after analyzing a paragraph written by that author. In this frame work we predict the trait of author under some specified condition. For making this system automatic we have used one of the supervised machines learning approach known as Support Vector Machine.The role of data set is to train the machine, after training the machine; we pass the test file on same machine, and it will return the accuracy. In our project we have used binary classification. The performance of project is directly proportional to the training data set. Larger the training data set more accuracy, This training set is generated human annotatable, and it is assumed that our training set is 100% accurate.

VI. FUTURE WORK

In this project we have used Support Vector Machine for categorization of the author's trait on the text. We will develop same frame work for the unsupervised learning approach e.g. HMM (Hidden Markov Model), SOP (Self Organized Map), etc. In this frame work we generate the feature vector matrix on eight features, and we decide the trait's under five categories. Beside these five categories there is some other categories e.g. Self-esteem, Harm avoidance, Impulsivity, Rigidity etc. Under these eight features results are very close to different trait. For making a good difference between all the classes it is necessary to take much more different features as much as possible e.g. Social words, Adjective related to personal status, Adjective related to achievements category, inclusion and exclusion words etc. Here we have applied binary classification with the help of Support Vector Machine. So there is another improvement area is classification. We apply multi-classification for classifying our text file. The Personality traits vary according to the adjective used by an author in their written text. We have a database from where we decide that a particular adjective belongs to which category. If the words exist in database there is no problem, but suppose words does not exist then automatically it will store in the database and user decide the category of that adjective. It is also an important improvement area that we use the dictionary or thesaurus in place of database.

REFERENCES

[1]. Brown, G. and Yule, G. Discourse Analysis. Cambridge: Cambridge University Press Buchanan, T. 2006.
[2]. Cohn, M., Mehl, M.R. and Pennebaker, J.W. Linguistic markers ofpsychological change surrounding Psychological Science, 15,687-693 2001.
[3]. Costa, P., and McCrae, R.R. Professional Manual. Psychological Assessment Resources, Odessa, FL1992.
[4]. Efimova, L., and de Moor, A. Beyond personal web publishing: Anexploratory study of conversational blogging practises. Proceedingsof the 37th Annual HICSS Conference. Big Island, Hawaii 2005.
[5]. Glance, N., Hurst, M., and Tomokiyo, T. BlogPulse: AutomatedTrend Discovery for Weblogs. in Proceedings of WWW New York,US 2004..
[6]. 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.
[7]. Alastair J. Gill, Jon Oberlander, and Elizabeth Austin. Rating e-mailpersonality at zero acquaintance. Personality and Individual Differences, 40:497–507 2006.
[8]. James W. Pennebaker and Laura King. Linguistic styles: Language use as an individual difference. Journal of Personality and Social Psychology, 77:1296–1312 1999.
[9]. S. Nowson and J. Oberlander. The identity of bloggers: Openness and gender in personal weblogs. AAAI Spring Symposium, Computational Approaches to Analyzing Weblogs, Stanford University, 2006.
[10]. I. H. Witten and E. Frank. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations.Morgan Kaufmann, 1999.
[11]. L. R. Goldberg. The structure of phenotypic personality traits.American Psychologist, 48(1):26–34, 1993.
[12]. Ying Li Ching Y. SuenTypeface Personality Traits and Their Design Characteristics DAS Boston, MA, USA June 9-11, 2010.
[13]. K. H.-Y. Lin, C. Yang and H.-H. Chen. What Emotions NewsArticles Trigger in Their Readers? Proceedings of SIGIR, 733-734, 2007.
[14]. Guiying Wei, Xuedong Gao, Sen Wu Study of text classificationmethods for data sets with huge features, 2nd International Conference on Industrial and Information Systems 2010.
[15]. Xiaojin Zhu. Semi-Supervised Learning Literature Survey,Computer Science, University of Wisconsin -Madison, 2008.
[16]. H Jia-wei, M Kamber. Data mining: concepts and techniques, 2ndedition, New York: Morgan Kaufmann Press, 2006.
[17]. AH-HWEE TAN, “Text Mining: The State of the Art and the Challenges”[C], PA KDD’99 Workshop on Knowledge Discovery from Advanced Databases（KDAD’99）, Beijing, 1999.
[18]. C. Sanderson and S. Guenter. Short text authorship attribution via sequence kernels, markov chains and author unmasking. In Proceeding of 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP), page 482491. Association for Computational Linguistics, 2006.
[19]. A. Abbasi and H. Chen. Applying authorship analysis toextremist-group web forum messages.IEEE INTELLIGENT SYSTEMS, pages 67–75, 2005.
[20]. 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.
[21]. M. Koppel and J. Schler. Exploiting stylistic idiosyncrasies for authorship attribution. In Proceedings of IJCAI’03 Workshop on Computational Approaches to Style Analysis and Synthesis, pages 69–72, 2003.
[22]. M. Gamon. Linguistic correlates of style: authorship classification with deep linguistic analysis features. In Proceedings of the 20th international conference on Computational Linguistics.Association for Computational Linguistics Morristown, NJ, USA, 2004.
[23]. ShlomoArgamon, Marin Saric, and Sterling S. Stein. Style mining of electronic messages for multiple authorship discrimination: first results. In Proceedings of SIGKDD, pages 475–480 2003.
[24]. ShlomoArgamon, SushantDhawle, Moshe Koppel, and James 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.
[25]. Satanjeev Banerjee and Ted Pedersen. The design, implementation, and use of the ngram statistics package. In Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics, pages 370–381, Mexico City 2003.
[26]. Tom Buchanan. Online implementation of an IPIP five factor personality inventory 2001. [27]. Paul T. Costa and Robert R. McCrae, Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI): Professional Manual. Odessa, FL: Psychological Assessment Resources 1992.
[28]. Francis Heylighen and Jean-Marc Dewaele. Variation in the contextuality of language: an empirical measure. Foundations of Science, Volume 7, pages 293–340, 2002.
[29]. Douglas Biber. Variation across Speech and Writing. Cambridge University Press, Cambridge, 1988.
[30]. Max Louwerse, Philip M. McCarthy, Danielle S. Mc- Namara and Arthur C. Graesser. Variation in language and cohesion across written and spoken registers. In Proceedings of the 26th Annual Conference of the Cognitive Science Society, pages 1035–1040, Hillsdale, NJ, 2004.
[31]. Scott Nowson, Jon Oberlander Differentiating Document Type and Author Personality from Linguistic Features Proceedings of the 11th Australasian Document Computing Symposium, Brisbane, Australia, December 11, 2006.
[32]. Chen, H., and Dumais, S. Bringing order to the web: automatically categorizing search results. CHI, 145-152, 2000.
[33]. Dumais, S., Chen, H. Hierarchical classification of Web content, In Proc. SIGIR, 256-263, 2000.
[34]. Yang, Y., Zhang, J., and Kisiel, B. A scalability analysis of classifiers in text categorization.SIGIR, 96-103, 2003.
[35]. Hersh, W., Buckley, C., Leone, T., and Hickam, D.OHSUMED: An interactive retrieval evaluation and new large test collection for research. SIGIR, 192-201, 1994.
[36]. Cai, L. and Hofmann, T. Hierarchical Document Categorization with Support Vector Machines, CIKM, 78-87, 2004.
[37]. Granitzer, M. Hierarchical text classification using methods from machine learning, Master's Thesis, Graz University of Technology, 2003.
[38]. Sun, A. and Lim, E. Hierarchical Text classification and evaluation, ICDM, 521-528, 2001.
[39]. Yang, Y., Zhang, J., and Kisiel, B. A scalability analysis of classifiers in text categorization.SIGIR, 96-103, 2003.
[40]. Allport, F. H. &Allport, G. W. Personality traits: heir Classification and Measurement. Journal of Abnormal and Social Psychology, 16, 1–40 1921.
[41]. Allport, G. W. Personality –A psychological interpretation. New York: Henry Holt and Company 1937.
[42]. Allport, G. W. The use of personal documents in psychological science. New York: Social Science Research Council 1941.
[43]. Allport, G. W. Becoming. Basic considerations for a psychology of personality. New Haven: Yale University Press 1955.
[44]. Allport, G. W. Letters from Jenny. New York: Harcourt, Brace & World, Inc 1965.
[45]. Allport, G. W. Traits revisited. American Psychologist, 21, 1-10 1966.
[46]. Cartwright, D. S. Introduction to personality. Chicago: Rand McNally College Publishing Company 1974.
[47]. Robert J. Harvey, William D. Murry, Steven E. Markham A “Big Five” Scoring System for the Myers-BriggsType Indicator Annual Conference of the Society for Industrial and Organizational Psychology Orlando in May 1995.
[48]. Barkhuus, L. and Csank, P Allport’s Theory of traits- a critical review of the theory and two studies A Technical report, Concordia University 1999.