Linguistic Features Based Personality Recognition Using Social Media Data

Dilini Sewwandi[1], Kusal Perera[2], Sajith Sandaruwan[2], Oshani Lakchani[1], Anupiya Nugaliyadde[1], and Samantha Thelijjagoda[3]
Department of Information Technology[1], Department of Software Engineering[2],
Department of Information Systems Engineering[3],
Sri Lanka Institute of Information Technology Malabe
samantha.t@sliit.lk

***Abstract* — Social media has become a prominent platform for opinions and thoughts. This stated that the characteristics of a person can be assessed through social media status updates. The purpose of this research article is to provide a web application in order to detect one's personality using linguistic feature analysis. The personality of a person has classified according to Eysenck's Three Factor personality model. The proposed technique is based on ontology based text classification, linguistic feature-vector matrix using LIWC (Linguistic Inquiry and Word Count) features including semantic analysis using supervised machine learning algorithms and questionnaire based personality detection. This is vital for HR management system when recruiting and promoting employees, R&D Psychologists can use the dynamic ontology for storage purposes and all the other API users including universities and sports clubs. According to the test results the proposed system is in an accuracy level of 91%, when tested with a real world personality detection questionnaire based application, and results demonstrate that the proposed technique can detect the personality of a person with considerable accuracy and a speed.**

***Keywords*— Ontology, Semantic Analysis, Eysenck's Three Factor model, Machine Learning Algorithms, LIWC (Linguistic Inquiry and Word Count) features.**

I. INTRODUCTION

In this world there are millions of people and every people is different with to each other with respect to individual features. This Research article targets in detecting the personality of a person by analyzing long term behavioral pattern of Facebook status and likes in a way to get faster and accurate decisions for HR (Human Resource) managing systems, university managing systems etc. Existing personality detection applications is only focus in LinkedIn or Twitter data and most of them are not using an ontology based system. When it comes to decision making process, employee profiling is a major component. Existing personality detection applications failed to provide efficient way of analyzing and managing backend of the system. When a company wish to buy this commercially available product, they need to buy personality detection system, dynamic ontology and dashboard separately which is a huge drawback. To bridge the gap, appropriate tools and technologies should be carefully selected focusing on the specific problem. This Research directs to the same approach by providing cost effective, efficient, ease to use system with personality detection system, dynamic ontology and dashboard as a single product. The special feature of our dashboard is it has fully customized and user can use it in any situation like HR (Human Resource) system, product search engine etc.

From the information carried out by previous Researches they have used self-assessment questionnaire based system to detect the personality of a person which induce a measurement error. In questionnaire based system the maximum word limit is 140 characters. In contrast to questionnaire based systems proposed system provides unlimited number of characters as the raw input to the system. A methodology to handle informal language are not provided by the existing systems other than proposed system [1]. Most of the existing systems are mainly focused in one aspect of detecting the personality such as machine learning or using LIWC features or using neural networks. But in this system machine learning techniques, extracting LIWC features and Ontology based personality detection are embedded together to handle most of the limitations that were emerged from the past researches.

II. BACKGROUND

A. Personality, Big Five Model and Eysenck's Three Factor Model

According to Gordon Allport "Personality is a dynamic organization, inside the person, of psychophysical systems that create the person's characteristic patterns of behavior, thoughts and feelings [2]."

According to Stephen Robins "Personality is the sum total ways in which an individual reacts to and interact with others.

Following describes the Big Five model (Five Factor Model) as measurement of personality. In Big-Five personality test explore your personality with the highly respected five factor model. The major dimensions of Big Five are

1. Extraversion vs. Introversion (sociable, assertive, playful vs. aloof, reserved, shy)
2. Emotional stability vs. Neuroticism (calm, unemotional vs. insecure, anxious)

3. Agreeableness vs. Disagreeable (friendly, cooperative vs. antagonistic, faultfinding)
4. Conscientiousness vs. Unconscientiously (self-disciplined, organized vs. inefficient, care-less)
5. Openness to experience (intellectual, insightful vs. shallow, unimaginative)

When targeting to the Eysenck's three factor model, the major dimensions are

1. Extraversion
2. Neuroticism
3. Psychoticism

1.) Focus to the Eysenck's Personality Theory

Each factor should describe using as a scale ranging from 'low' to 'high', with a score possible at either extreme or anywhere in between. These factors are considered to be orthogonal and independent of each other, and therefore if an individual has a particular score on one factor, this does not necessarily predict any of their scores on any of the other factors. EPI (Eysenck Personality Inventory; Eysenck and Eysenck, 1964) which solely measures these traits.

2.) Mapping the Eysenck's three factor model, with Big Five model (Five Factor Model)

In here Openness is something similar to Extraversion. Agreeableness and Conscientiousness is placed under Psychoticism. Agreeableness and Conscientiousness is inversely proportional to the Psychoticism [3], [4], [5], [6].

B. Semantic Analysis

In linguistics, semantic analysis is the process of relating syntactic structures, from the levels of phrases, clauses, sentences and paragraphs to the level of the writing as a whole, to their language-independent meanings.

C. LIWC (Linguistic Inquiry and Word Count)

LIWC is a transparent text analysis program that helps to counts words in psychologically meaningful categories. Empirical results using LIWC depicts its ability to detect meaning in a broad variety of experimental settings such as showing attention focus, emotionality, social relationships, thinking styles, and individual differences [7].

D. Naïve Bayesian

Equation (1) is Bayes theorem is used to calculating the posterior probability, $P(c|x)$, from $P(c)$, $P(x)$, and $P(x|c)$. Naive Bayes classifier assume that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors. This assumption is called class conditional independence [8].

$$P(C/X) = (P(X/C)\,P(C))/\,(P(X)) \quad \text{——————} (1)$$

$$P(C/X) = P(X_1/C)*P(X_2/C)*\ldots*P(X_n/C)*P(C) \quad \text{———} (2)$$

Explanation of Equation (1) and (2)

1. *P (c|x) is the posterior probability of class (target) given predictor (attribute).*
2. *P(c) is the prior probability of class.*
3. *P (x|c) is the likelihood which is the probability of predictor given class.*
4. *P(x) is the prior probability of predictor.*

E. Ontology

Ontology is collection of classes and sub classes which are inter related and properties are describing the features and attributes of the classes and restrictions. Ontology with instances for classes creates a Knowledge Base [9]. These classes denote entities of specific domain and interrelation in between these classes [10].

Ontologies have widely used in knowledge sharing and management system recently. Most of the expertise Systems are created based on an Ontology.

F. Use of an Ontology

The reasons to use ontology is to share common understating of the information structure, reuse and analyze the domain knowledge [9].

III. METHODOLOGY

Prototype methodology has followed in order to carry out this research project. As a pre-preparation for the Research, a broad investigation was carried out to identify major business problems faced by the companies. There weren't any existing applications to detect one's personality in the aspect of business development such as recruiting employees, granting allowances and incentives and granting promotions. Proposed system is the solution for this drawback.

A traditional feasibility study was followed in order to check whether the proposed system is financially, technically, and operationally feasible. In there is has proved that due to the open source technologies and doesn't have any technological constraints and dependencies. Then the motivation was directed to functional and non-functional approach of the system.

In design phase, gathered functional and non-functional requirements of this system was mapped to a high-level architecture.

In the implementation phase proposed system has focused to detecting the personality of a person in three dimensions.

A. Ontology based personality detection.
B. Personality detection through linguistic analysis.
C. Questionnaire based personality detection
D. Before directing to the personality detecting dimensions the main focus was extracting publicly available data from the Facebook in an ethical way.

The high level architectural diagram of the system has illustrated in Figure 1 below, which consists all three personality recognition dimensions mentioned above and how the data extraction process has incorporated in all three mechanisms.

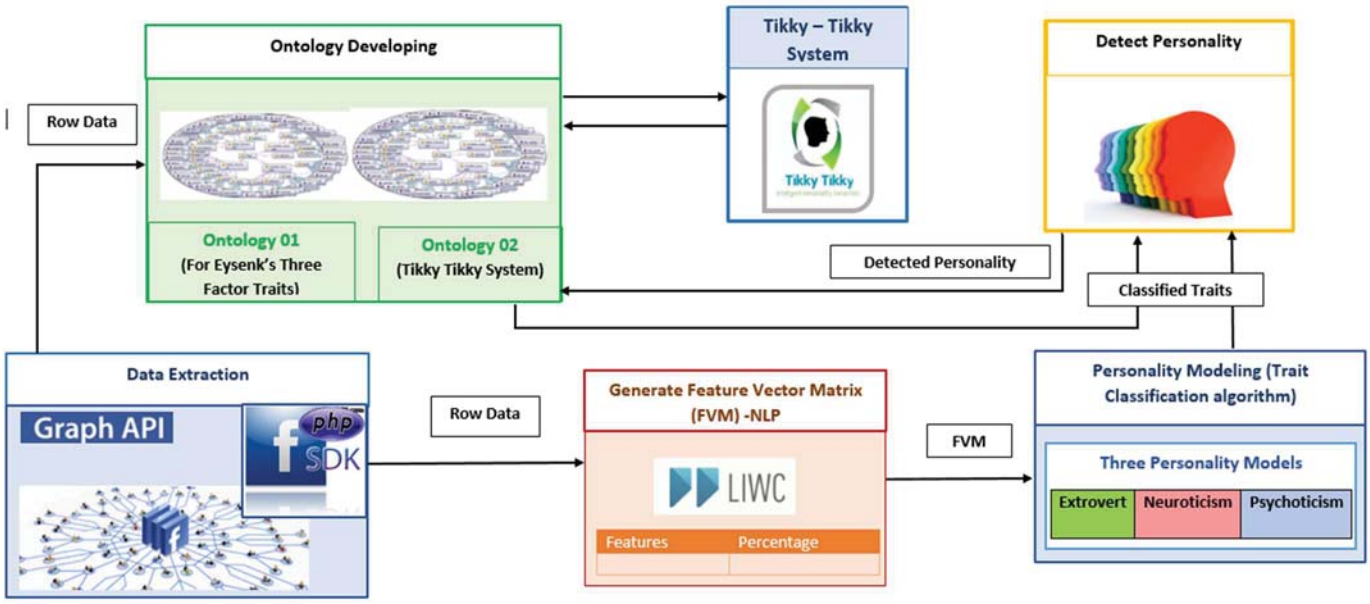

Fig. 1. High Level Architectural Diagram of the Proposed System

API which is a restful API the Facebook data was extracted through PHP SDK. By default, Facebook is not allowed to extract the publicly available data of a particular user without user permission. Therefore, this web application has integrated with a Facebook application and user has by default access to the Facebook application when the Facebook user credentials has passed to the web application and through that user will be authenticated. When the user has logged to the system in the very first time this web application will be requested some permissions such as user_friend, user_feed and user_likes etc. Then only data can be extracted for further tasks of the proposed system.

A. Ontology based personality detection.

The main approach of detecting the personality of proposed system is Ontology based personality detection. In here the ontology was designed using OWL ontology developing language using protégé software with the aid of OWL-DL package.

Use Protégé OWL

- Protégé OWL provides multiuser support for synchronous knowledge entry.
- Protégé OWL can be extended with back-ends for alternative file formats. Current formats include Clips, XML, RDF, and OWL.

The ontology was created using a data set acquired by a well-known Psychologist in Sri Lanka based on the above mentioned Eysenck's three factor model. Categories in the ontology has divided under main categories, sub categories and individual levels. The words are entered to the leaf levels. An example is shown in Figure 2, considering the main category as Neuroticism, two sub categories called moody and anxiety that are related to the Neuroticism and the individuals that are related to those subcategories are displayed.

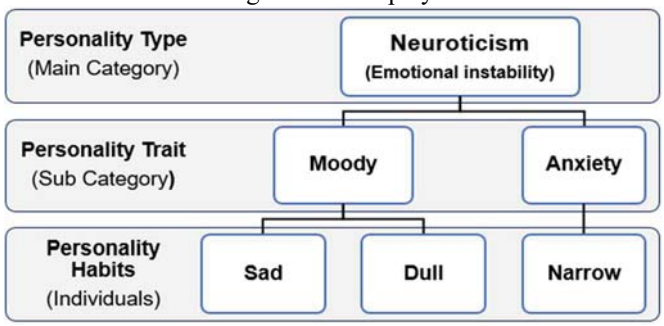

Fig. 2. Ontology Development Hierarchy

When querying the ontology with the aid of the Jena query engine an algorithm was developed to retrieve data from the ontology.

For an example assume that the word sad has entered to the system, then hierarchy will be shown that it is belongs to moody sub category under Neuroticism main category.

In here as the input to the system Facebook status are provided. After doing tokenization, lemmatization and re-correcting of spelling mistakes for those status and select the words that are in the ontology only and discard other personality in-related words. At the final stage of the text processing in ontology acquire the main categories and sub categories percentages to detect the personality as in Figure 3.

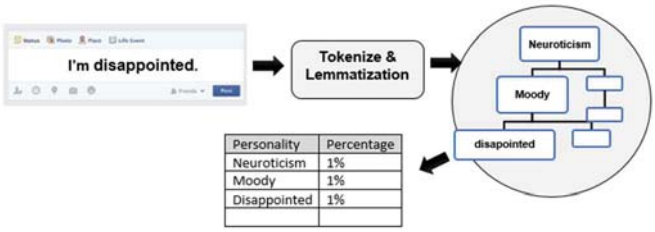

Fig. 3. Personality detection through an ontology

B. PERSONALITY DETECTION THROUGH LINGUISTIC ANALYSIS.

Linguistic based personality detection was done based on LIWC features. From the extracted word set, the focus was totally driven to get the percentage of those features in the inputted statuses [7].

Some of the main features used for the linguistic analysis is shown below.

1. Large word percentage
2. Percentage of personal pronouns.
3. Lowercase percentage.
4. Percentage of prepositions.
5. Percentage of 1st person singular pronouns.
6. Percentage of pronouns.
7. Percentage of 3rd person pronouns.
8. Percentage of present tense verbs.
9. Percentage of short sentences.
10. Small words percentage.
11. Uppercase percentage.
12. Percentage of nouns.
13. Percentage of past tense verbs.
14. Percentage of articles.
15. Percentage of 1st person plural pronouns.
16. Percentage of long sentences.
17. Percentage of adverbs.
18. Percentage of positive emotional sentences.
19. Percentage of negative emotional sentences.
20. Percentage of words that have spelling mistakes.

Facebook statuses were given as the input when detecting the personality through linguistic analysis as well. Python NLTK library was used for natural language processing tasks. To find several feature percentages tokenization, stemming and lemmatization was done before directed to the feature vector matrix. In order to identify negative and positive emotions six machine learning algorithms have used. They are original Naive Bayes Algorithm [11], Multinomial Naïve Bayes Classifier, Bernoulli Naïve Bayes Classifier, Logistic Regression Classifier, Linear Support Vector Machine Classifier and Stochastic Gradient Descent Classifier. From all these algorithms the sentiments are detected from the best algorithm. After that percentage of positive emotions and negative emotions have calculated as a dimension of feature vector matrix. After getting those percentages initial step of building the scoring model was done according to the linguistic hypothesis. Using the experimental results used in [12] which was already developed for Extroversion and Introversion boundaries, further extended up to Neuroticism, Emotional Stability and Psychoticism, Tender as well. The hypothesis that was needed to build the scoring model has summarized according to all the personality traits [16], [17], [18]. As an example in TABLE I, the hypothesis required for Extrovert category has shown [1], [14], [15].

TABLE I. SUMMARY OF LINGUISTIC HYPOTHESIS FOR EXTRAVERT

	Realisation	Grammatical	LIWC
Extraversion	Loudness ✓ Capital Letters ✓ Exclamation marks Worse Pronunciation ✓ Worse spelling ✓ Typographical error	✓ Adverbs ✓ Pronouns ✓ Verbs ✓ Lexical Density ✓ Nouns ✓ Prepositions ✓ Formal	✓ Social and positive emotion words ✓ Negations ✓ Tentatively ✓ Exclusive ✓ Inclusive ✓ Causation ✓ Negative emotion words ✓ Articles

C. QUESTIONNAIRE BASED PERSONALITY DETECTION

The targeted sector for questionnaire based mechanism is the users who does not have a Facebook account. They can manually log in to the system and answer the questions in order to gain the graphical representation of their personality based on Big Five model.

In here as in [13] 50 questions were used in order to detect the personality based on Big Five personality traits. The sentences were categorized under those five traits and each and every sentence has given a weightage. In answering options for + keyed items, if the response "Very Inaccurate" is assigned a value of 1, "Moderately Inaccurate" a value of 2, "Neither Inaccurate nor Accurate" a 3, "Moderately Accurate" a 4, and "Very Accurate" a value of 5. For - keyed items, if the response "Very Inaccurate" is assigned a value of 5, "Moderately Inaccurate" a value of 4, "Neither Inaccurate nor Accurate" a 3, "Moderately Accurate" a 2, and "Very Accurate" a value of 1. Finally, the personality will be represented using a bar graph based on Big Five model.

1.) Analyzing Facebook Likes
   - Graph API allows to extract categories of the pages that user have liked that encourages to identify the users likes and dislikes as well.

2.) Dashboard of the system (Admin-End)
   - Dashboard is the web-based component in the system which provides fully customized basic admin functionalities mainly focused to a HR (Human Resource) system, but can be extensible for others as well. Some of the major functionalities in admin end are listed below.
     1. Positions Management.

2. Department Management.
3. Employee Personality Management.
4. Ontology Management.
5. Question Management.
6. Personality Attributes Management.

IV. RESULTS AND DISCUSSION

Intended for testing the reliability of any constructed system a test data is mandatory. Tested the proposed system with different data souses and set-ups such as Carl Jung and Isabel Briggs Myer's personality type theory and testing tool.

System testing approach will show in the TABLE II. System testing starting with of three main categories students' groups such as SLIIT first year students, SLIIT Curtin students and A/L students. For each testing group consist of ten students. A questionnaire and a personal interview has conducted to manually categorize them into main six trait hypothesis models.

Based on their response to the questions in the questionnaire all experimental results are given in TABLE II. Then the personality of the students belongs to different samples have tested with the system and map all the experimental data into the below table and used them to test the accuracy of the proposed personality detection system.

TABLE II. EXPERIMENTAL RESULT TABLE
FOR DIFFERENT STUDENT CATEGORIES

Category	Student Name	Extrovert	Introvert	Neuroticism	Emotional Stability	Psychoticism	Tender
SLIIT 1st Year Students	Test01	Agree	Agree	Agree	Agree	Agree	Agree
	Test02	Disagree	Disagree	Agree	Agree	Agree	Agree
	Test03	Agree	Agree	Disagree	Disagree	Agree	Agree
	Test04	Disagree	Disagree	Agree	Agree	Disagree	Disagree
	Test05	Agree	Agree	Agree	Agree	Disagree	Disagree
SLIIT Curtin Students	Test06	Agree	Agree	Agree	Agree	Disagree	Disagree
	Test07	Agree	Agree	Agree	Agree	Agree	Agree
	Test08	Disagree	Disagree	Disagree	Disagree	Agree	Agree
	Test09	Agree	Agree	Agree	Agree	Disagree	Disagree
	Test10	Disagree	Disagree	Agree	Agree	Agree	Agree
A/L Students	Test11	Agree	Agree	Agree	Agree	Disagree	Disagree
	Test12	Disagree	Disagree	Agree	Agree	Disagree	Disagree
	Test13	Disagree	Disagree	Disagree	Disagree	Agree	Agree
	Test14	Agree	Agree	Disagree	Disagree	Agree	Agree
	Test15	Agree	Agree	Agree	Agree	Agree	Agree

Following assumptions have been arranged in the working set of proposed system.

1. Assumed that the text to be attached to any word.
2. Full stop must not be attached to any word.
3. Unlimited extracted data can be test.

E.

After the multiple runs on above experimental data, the average accuracy of the proposed system has been calculated as 91%. Final system was using Facebook statuses and Likes. After summarized the approaches reviewed in the previous research articles [1], [14], [15], [16], [17], [18] and how proposed system has come up with the new solutions to those limitations. It can be inferred that most of the approaches typically require the users to fill a form containing several questions and then use this inventory to predict personality based on big five model. But in this approach personality is detecting through three dimensions as mentioned before. In the admin end user can edit main categories, sub categories, leaf level objects and data properties of the ontology as the wish of the user because of the generic algorithm that was used to develop this system.

V. CONCLUSION

Social media has opened the door to share information, attitudes and as well as get feedbacks from others as well. The revelation of views of individuals inspires other users to comments and pay their attitude towards others views as well. This may expose to detect the personality that beneficial for many areas including HR management, psychology, medical and business intelligence as well. Using personality traits based on Big Five model and Eysenck's Three Factor model, a description of a person can be given by analyzing the linguistic usage of that person. This research paper provided some techniques that can be used to detect the personality of a person by ontology based personality detection, with an accuracy of around 91% when tested with a real world questionnaire based application. System will be exposed as an API for universities, sports and social clubs when recruiting individuals to those organizations. This research article has discussed some strengths and limitations of these approaches. The conclusion of this research article is in order to improve detection capacity the consideration of texts as well as social behavioral aspects of a user on multiple social media (e.g. Twitter, Facebook, and LinkedIn) it would be more beneficial.

VI. ACKNOWLEDGEMENT

We would like to acknowledge the efforts of consultant psychiatrist Dr. Rumi Ruben in the data collection and guidance.

REFERENCES

[1] A. B. R. B. a. G. J. P. C. Sumner, "Predicting Dark Triad Personality Traits from Twitter usage and a linguistic analysis of," in *IEEE 11th International Conference on Machine Learning and Applications ICMLA*, 2012.

[2] S. A. L. a. M. Windelband, "The Person as a Focus for Research –," *Journal for Person-Oriented Research,* 2015.

[3] A. J. Gill, Personality and language : the projection and perception of personality in computer-mediated communication, 2004.

[4] A. J. G. a. J. O. Gill and Oberlander, "Taking care of the linguistic features of," in *4th Annual Conference of the Cognitive Science Society (CogSci2002)*, Fairfax, VA, USA., 2002.

[5] J. O. a. A. Gill, "Language generation and personality: two dimensions, two stages, two hemispheres?," *American Association for Artificial Intelligence,* 2004.

[6] J. a. G. A. Oberlander, "Individual differences and implicit language: personality, parts-of-speech and pervasiveness," *26th Annual Conference of the Cognitive Science Society,* vol. 1040, p. pp1035, 2004.

[7] J. W. P. Yla R. Tausczik, "The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods," *Journal of Language and Social Psychology,* 2010.

[8] D. S. Sayad, "An Introduction to Data Mining," 2010. [Online]. Available: http://www.saedsayad.com/data_mining_map.htm. [Accessed 20 March 2016].

[9] N. F. N. a. D. L. McGuinness, "Ontology Development 101: A Guide to Creating Your First Ontology," National Institute of General Medical Sciences, 2016 . [Online]. Available: http://protege.stanford.edu/publications/ontology_development/ontology101-noy-mcguinness.html. [Accessed 15 March 2016].

[10] P. S. a. N. V. Pankajdeep Kaur, "An Ontology Based Text Analytics on Social Media," *International Journal of Database Theory and Application,* vol. Vol.8, p. No.5, 2015.

[11] A. K. Kanupriya Sharma, "A review of the existing state of Personality prediction of Twitter," *IOSR Journal of Computer Engineering (IOSR-JCE),* vol. Volume 17, no. Issue 4, pp. p-ISSN: 2278-8727, 2015.

[12] C. M. S. Saurabh Saxena, "Machine Learning based approach for Human Trait Identification from Blog Data," *International Journal of Computer Applications (0975 – 888),* vol. Volume 48, p. No.10, 2012.

[13] U. N. I. o. M. Health, "International Personality Item Pool (IPIP)," 03 November 2016. [Online]. Available: http://ipip.ori.org/. [Accessed 15 June 2016].

[14] C. R. M. E. a. K. T. Jennifer Golbeck, "Predicting Personality from Twitter," in *2011 IEEE International Conference on Privacy, Security, Risk, and Trust, and IEEE International Conference on Social Computing*, 2011 .

[15] A. G. A. C. H. Soujanya Poria, "Advances in Soft Computing and Its Applications, Publisher: Common Sense Knowledge Based Personality Recognition from Text," in *12th Mexican International Conference on Artificial Intelligence*, 2013.

[16] W. D. Ben Verhoeven, "Ensemble Methods for," in *Association for the Advancement of Artificial*, 2013.

[17] T. K. C. S. Randall Wald, "Machine prediction of personality from Facebook profiles," in *Information Reuse and Integration (IRI), 2012 IEEE 13th International Conference on*, Las Vegas, 2012.

[18] IEEE, "Our Twitter Profiles, Our Selves: Predicting Personality with Twitter," in *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference*, Boston, 2011.