ISSN: 0976-3031

Available Online at http://www.recentscientific.com

International Journal of Recent Scientific Research
Vol. 8, Issue, 1, pp. 15434-15438, January, 2017

International Journal of Recent Scientific Research

Research Article

PERSONALITY TRAITS DETECTION USING RSTUDIO

Shefali Sharma[1]*., Rujutha Shetty[2]., Bini Shah[3] and Seema Yadav[4]

[1,2,3,4]Department of Information Technology, K.J Somaiya Institute of Engineering and Information Technology, University of Mumbai, Maharashtra

ARTICLE INFO

Article History:

Received 18th October, 2016
Received in revised form 10th November, 2016
Accepted 06th December, 2016
Published online 28th January, 2017

Key Words:

Data mining, Data source, Personality traits, Criminal Psychology, Classification, Clustering

ABSTRACT

Increasing crime rates all over the world have lead us to some serious concerns. While we are busy finding punishments for criminals, we tend to forget the fact that prevention is always a better solution. Human behavior is strongly linked to the environment we thrive in. In today's world, social media has become the key to deciphering a personality. Through our system, we aim to utilize the data extracted from various social media platforms in order to narrow down the key factors which might lead to a potential criminal mindset. We make use of a premium integrated development environment i.e RStudio to arrive at a conclusion which will be in terms of criminal potential in any individual. Personality trait analysis has been a subject of interest for years but in the proposed system, we aim to use clustering as well as classification algorithms to have a more accurate analysis of the personalities which will be classified on the basis of the Five Factor Model. The five factors have been defined as openness, conscientiousness, extraversion, agreeableness and neuroticism (OCEAN).

INTRODUCTION

Data mining [Yun Xiong, Yangyong Zhu, 2009] is the computational process used for discovering a number of patterns in large data sets involving techniques at the crossing points of machine learning and statistics, database systems and artificial intelligence. The main goal of data mining process is to obtain information from an existing data set and transform it into a simplified structure for further utilization.

The idea of our proposed system evolved because of the increased crime rates in the society. We are designing the system in such a way that it will help to decrease the crime rate. As most of the population is actively involved in the social media sites, we can get an abstract view of the mindset of the particular suspect.

Thus, this system proves to be a boon for the police department to have a broader approach in solving the case. It is extremely important to mine the peculiar objects' groups as this will help in the criminal background check. Personality analysis based on this mining also plays a critical role as it has a lot of potential in practice.

Mining peculiarity groups and defining a measurement of the degree of peculiarity is a major concern. When it comes to correlating the same to personality trait analysis, these peculiarity groups can be defined in terms of the Five Factor Model [Menasha Thilakaratne, Ruvan Weerasinghe, Sujan Perera, 2016]. It is a standard in Psychology with the help of which we can define the attributes of an individual. Extraversion, Neuroticism, Agreeableness, Conscientiousness and Openness to New Experiences are the most important ones.

The usage of a certain set of keywords or continuously choosing to visit some particular site shows an ardent inclination towards that respective personality trait. Here, we are designing the system in such a manner that the data extracted from various social media platforms which are ripe with semantically rich information and also knowledge related to public domains gives us an insight into the inclination of the individual towards a particular arena of information. This will help us to show the direct and indirect association to a particular type of trait. This can be done by efficiently classifying the keywords into clusters which will denote the traits. The further analysis on these proportions of personality traits will reveal the potential that a person might be a criminal or not. The existing systems have provided a generalized view of the personalities present in a person but our proposed system moves a step ahead by detecting the criminal behavior in that particular individual using their social media data. Also the proposed system makes use of the latest technology RStudio, which has never been used earlier for personality trait analysis. This tool provides the results that are very easy to interpret unlike many other complex tools.

*Corresponding author: **Shefali Sharma***
Department of Information Technology, K.J Somaiya Institute of Engineering and Information technology, University of Mumbai, Maharashtra

Existing systems

Twitter data set personality

The user generated content on twitter i,e tweets also provides a source of information for identifying users personality traits. It was considered that only few users posted links on twitter, which formed the content of the dataset. This dataset has been used for the task of automatically predicting the personalities of the users, as well as for user behavior analysis. It was observed that extroverts and emotionally sound people are popular as well as users who are influential on twitter[8]. It Was Also Observed That Popular Users Are Creative, While Influential People on Twitter Are More Organized.

Using Clustering To Extract Personality Information From Socio Economic Data

In this work, a method to extract behavior related to different groups by using simple clustering methods that can potentially reveal various aspects of the personalities [Alexandros Ladas, Uwe Aickelin, Jon Garibaldi, Eamonn Ferguson, 2013]. Extraction of different clusters such as selfish and non – selfish behavior is based on the judgments of 52 students from psychology background that rated the 9 attributes of expenditure of the dataset. The results demonstrated that it is possible to extract information regarding the personality of individuals from similar datasets by using even simple data mining techniques.

Improving User Profile with Personality Traits Predicted From Social Media Content

Here, the personality was proved to be related with personal preferences in music tastes and guis. The systems that recommend accuracy should be improved by taking user personality traits into consideration. Results of these survey Indicated That Personality Could affect the Desire of online purchasing activity. Improving user Profile with Personality Traits Could Enhance Performance of The Recommendation System [Gao Rui, Hao Bibo, Bai Shuotian, Li Lin, Li Ang, Zhu Tingshao,2013]. The Recommendation Systems Which Are Personalized Based on Social Media Usage Behavior And Content is among the apex topics throughout the whole network in both academics and business. However, the personality traits of user were mainly measured with psychological questionnaire.

Table 1 Comparision Proposed design (Proposed System and Existing System)

Parameters	Proposed System	Existing System
Personality Detection	Basic personalities present in an individual along with analysis of criminal psychology will approximate the chances of a person likely to be a criminal.	Personality traits of a person at a very basic level without further analysis have been implemented.
Accuracy	Both Classification and Clustering algorithms are used which increases the accuracy.	Classification and Clustering are not applied together in the existing systems.
Platform	Most commonly used messaging application i.e Whatsapp is used.	Facebook as well as Twitter data have been used for analysis.
Technology Used	R Programming language will be used due to it's special feature of statistical computing along with analysis.	Implementation is done using most commonly used programming language.

It is difficult to obtain personality traits of large amount of users in real-world personalized systems in case of application scenes. Thus the system provides an efficient and approximate result of The Personality Analysis of An Individual.

Comparision

Proposed Design

The Proposed Design Is Divided Into Five Modules

Data Source

The data source that we will mainly focus on is whatsapp chats as well as twitter tweets. Whatsapp chats will be extracted using the 'email chat' option. Twitter data will also be extracted by direct authentication method using twitter application. The data obtained from these sources will be obtained in unstructured data.

Data Extraction

The data will be extracted from the data sources using r, which is a programming language and environment for statistical computing and graphics. Rstudio is a free and open-source ide for r. Data has been extracted from whatsapp and twitter using the corresponding commands. Whatsapp data was extracted via the email. Similarly, twitter data was extracted by creating a twitter application and generating the consumer key, consumer access secret key and so on**.**

R And Mysql Connection

The retrieval and storage of data from a mysql database with r is possible by using the rmysql package. The package simply needs to be installed and loaded in the library of the database. Also we can directly load the data into rstudio by installing the twitter package. Thus the tweets are obtained in a cleaned format.

Classification and Clustering Algorithms

Classification and clustering algorithms will be applied on the data to identify the personality traits of a person based on various criteria considering the base as the ‘five factor model of personality’. The big five traits of a human are openness, conscientiousness, extraversion, agreeableness and neuroticism. Naive bayes' theorem will be used for the classification of the personalities of that individual. There are many in-build commands for the same.

User Interface

A user interface may be used to display the results in an appropriate manner in the form of statistical diagrammatical representations such as graphs and pie charts. Analyzing the conclusion, we will display the chances of a person likely to be a criminal. The details of the graph can be put in the rstudio to get the histogram. This will help us to show the association to a particular type of trait in Terms of Percentage and Other Parameters.

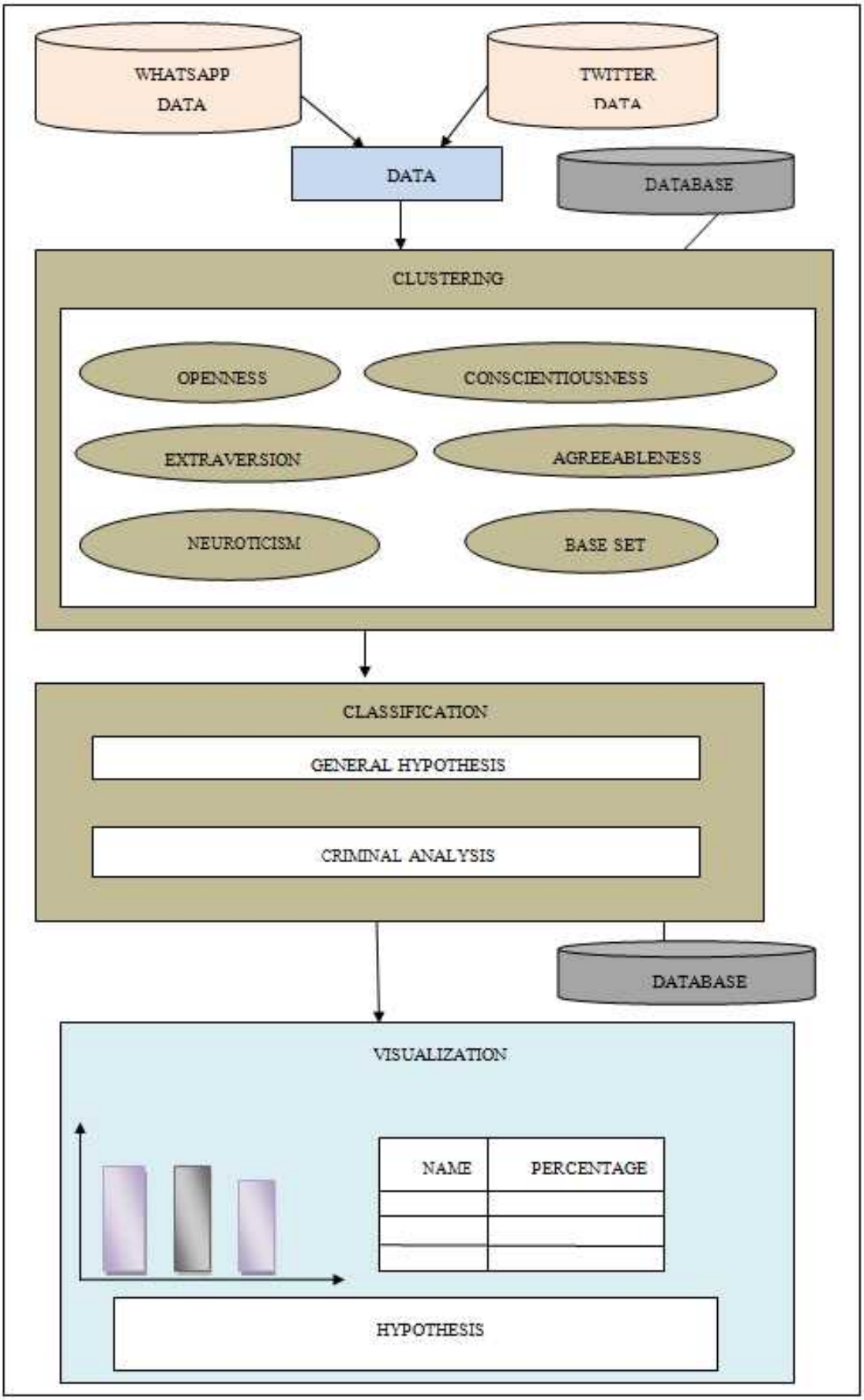

Figure 1 Architecture of our System

The Phases of Our Proposed System Is As Follows

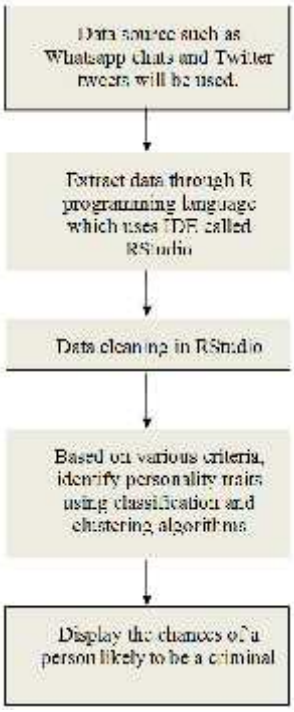

Figure 2 System Flow Diagram of Our System

Algorithms

Naive Bayes classifier is an algorithm which forms a set of supervised learning algorithms completely based on applying the Bayes theorem with an inappropriate assumption of independence which exists between every pair of characteristic features. Bayes' theorem will state the following: P(H|X) = P(X|H)P(H)/P(X)

where, P(H|X) is the posterior probability of H conditioned on X
P(H) is the prior probability of H
P(X) is the prior probability of X

Naive Bayes classification with e1071 package

The e1071 package consists a function named naiveBayes() which helps in performing Bayes classification. The function is able to receive categorical data and continguous table as input data. It returns an object of class "naiveBayes". The object can be passed to predict() to predict the outcomes of unlabeled subjects.

Naive Bayes classification with caret package

The package of caret contains train() function which helps in setting a grid of tuning parameters for many classification and routines of regression, fits each model and calculates a resampling based performance measure.

Clustering

R has an amazing variety of functions for cluster analysis. K Means Clustering is an unsupervised learning algorithm that tries to cluster the data based on their similarity. Unsupervised learning means that there is no result to be predicted and the algorithm just tries to find patterns in the data. In Kmeans clustering, we specify the number of clusters we want the data to be grouped into. The algorithm might randomly assigns each observation to a cluster, and finds the centroid of each cluster. Then, the algorithm iterates through two steps:

- Reassign various data points to the cluster whose centroid is near.
- Calculate the new centroid of each cluster.

These two steps are repeated until the cluster variation cannot be reduced further. The cluster variation is calculateed as the average sum of the euclidean distance between the data and their respective cluster centroids.

Advantages

When it comes to analyzing personalities, a lot of methods have already been applied. However, the efficiency in each of these models has been very less compared to other methods. Considering our system, we have used clustering as well as classification algorithms which will help us deduce the conclusions for this analysis. Initially the clustering algorithm provides the clustered version of the personality traits based on the five factors and its corresponding keywords. Now the clustered data makes use of the class labels property to depict the criminal habits with the help of the 'Neurotisim' keyword. We make use of the R programming language which is the most comprehensive statistical analysis package available. It incorporates most of the statistical tests, models and analyses, as well as provides a comprehensive language for managing and manipulating data.R plays well with many other tools, importing data, for example, from CSV or directly from Microsoft Excel, Microsoft Access, MySQL, and SQLite. It also produces graphics output in PDF, JPG, PNG, and SVG formats.

Applications

Criminal background check

Whenever there is an occurrence of any criminal activity and the police department decides to investigate the case, the police department will do a background check to determine the mindset of the criminal by gathering information from certain people. However, our system can be used to identify the mindset of the concerned person by accessing his/her social media accounts that would determine the personality based on the five factors which in turn would determine whether the person has a criminal mindset. This would play a major role in solving the case.

Medical Data Analysis

Medical area [K.M. Al-Aidaroos, A.A. Bakar and Z.Othman, 2012] produces extremely humongous and voluminous quantities of electronic data that is becoming more and more complicated. The generated medical data has certain characteristics which makes their analysis highly challenging and lucrative. In this method of Naive Bayes utilization for classifying medical data, it allows mining from different perspectives; including attributes of medical data, requirements of designs and systems dealing with this data and also the different methodologies used for medical data mining.

CONCLUSION

In this paper we have presented a way to analyze information which is obtained from extraction of real time data. The data

will be used to analyze the criminal psychology of an individual based on the Five Factor Model of Personality [Yun Xiong, Yangyong Zhu, 2009] i.e agreeableness, openness, extraversion, neuroticism, conscientiousness.

Our findings demonstrate that it is possible to determine the personality of individuals by using clustering as well as classification algorithm. This paper represents a way to make use of personality traits obtained from social media in order to determine criminal behavior amongst users. The scope of our proposed system extends to the field of cryptography. The encrypted data from the social media can be extracted and further analyzed from various aspects by the ethical hackers.

References

1. Menasha Thilakaratne; Ruvan Weerasinghe; Sujan Perera, "Knowledge-Driven Approach To Predict Personality Traits By Leveraging Social Media Data", 2016, P288-295, DOI 10.1109/WI.2016.47. Publisher: IEEE.
2. Yun Xiong; Yangyong Zhu," Mining Peculiarity Groups In Day-By-Day Behavioral Datasets", 2009 Ninth IEEE International Conference on Data Mining , 2009, p578-587, 10p. Publisher: IEEE.
3. Lima, Ana C.E.S.; castro, Leandro N," Multi-label semi-Supervised Classification Applied to Personality Prediction in Tweets" De. 2013 Brics Congress on Computational Intelligence & 11th Brazilian Congress on Computational Intelligence, 2013, P195-203, 9P. Publisher :IEEE
4. K.M. AL-Aidaroos, A.A. Bakar And Z.Othman, "Medical Data Classification With Naive Bayes Approach", Information Technology Journal 11(9) 1166-1174, 2012, ISSN: 1812-5638/DOI: 10.3923/ITJ. 2012.1166. 1174
5. XI Chen, Hemant Ishwaran, "Random Forests For Genomic Data Analysis", Us National Library Of Medicine National Institue of Health, Pubmed, MID:22546560, PMCID: PMC3387489, DOI:10.1016/J. YGENO.2012.04.003
6. Alexandros Ladas, Uwe Aickelin, Jon Garibaldi, Eamonn Ferguson, "Using Clustering To Extract Personality Information From Socio Economic Data", UKCI 2012, The 12th Annual Workshop on Computational Intelligence, Heriot-Watt University, 2012,Submitted On 8 Jul 2013
7. Gao, Rui; Hao, Bibo; BAI, Shuotian; Li, Lin; LI, ANG; Zhu, Tingshao, "Improving User Profile With Personality Traits Predicted From Social Media Content", Proceedings Of The 7th Acm Conference Recommender Systems, 10/12/2013, P355-358, 4P. Publisher: Association for Computing Machinery.
8. Farnadi, Golnoosh; Sitaraman, Geetha; Sushmita, Shanu; Celli, Fabio; Kosinski, Michal; Stillwell, David; Davalos, Sergio; Moens, Marie-Francine; Cock, Martine," Computational Personality Recognition In Social Media", In: User Modeling And User-Adapted Interaction. June 2016, VOL. 26 Issue 2-3, P109, 34 P.; Springer Language: English, Database: Academic Onefile.
9. W.A. Awad And S.M. Elseuofi, "Machine Learning Methods FOR SPAM E-MAIL Classification", *International Journal of Computer Science &Amp; Information Technology* (IJCSIT), VOL 3, NO 1,FEB 2011, DOI : 10.5121/IJCSIT.2011.3112 173
10. Ping Sun, Irena Begaj, Iris Fermin, Jim Mcmanus" Creating Health Typologies With Random Forest Clustering,"The2010 International Joint Conference on Neural Networks, Ieee Conference Publications, YEAR:2010, PG(1-7), DOI:10.1109/IJCNN.2010.55965

****

How to cite this article:

Shefali Sharma *et al*.2017, Personality Traits Detection Using Rstudio. *Int J Recent Sci Res.* 8(1), pp. 15434-15438.