Identifying Personality Traits using Social Media

MAKKENA PRAMOD[1], MIKKILI RAJ KUMAR[2], PAMULA ANIL KUMAR[3], NIZAMPATNAM NAGA SARATH[4], K. VIKAS[5]

[1,2,3,4] *Students, Computer Science Department, Vasireddy Venkatadri Institute of technology, Andhra Pradesh, India*

[5]*Asst. Professor, Computer Science Department, Vasireddy Venkatadri Institute of technology, Andhra Pradesh, India*

Abstract -- The Social media is no more another idea today. With the expansion in the infiltration of web and minimal effort cell phones access to online networking has more turned into a pattern and need too many. Having more number of preferences and plenty of remarks and further sharing the posts has turned into a societal position and notoriety issue to the adolescent of today. The special part of these highlights is that likes, comments, and shares are moment reactions of users and it is freely accessible and can be gotten to by every one of the companions of a man. Thus, Facebook is a popular social media platform with millions of users. The suggested framework takes the likes, comments and shares as input and processes the same to map it to a personality trait. The paper derives the framework by considering big five personality traits.

Keywords— Data Mining; Personality Traits; Social media; Text analytics.

I. INTRODUCTION

Social media has become an essential component of everyday life. It has radically changed the ways in which people express their opinions and sentiments. Social networking sites like Facebook, LinkedIn and Twitter are based on the concept of human interactions and user generated data. Thus a huge amount of user data is been created and exchanged, entailing massive production of interactive data. Social media-oriented people tend to publish a lot about themselves through status updates, self-description, photos, videos and interests. The data available within social media platform is enormous in volume and reveals different aspects of human behavior and social interactions. An individual's personality is his/her characteristics and aspects that others perceive. The data available on social media platform enables us to understand who the users are and what their needs are. Thus, the analysis of social media data allows us to determine important personality traits i.e., characteristics which describes his/her personality.

Personality is a way person respond to a particular situation. It is combination of characteristics that make an individual unique. Assessment of personality over the past two decades in various researches has revealed that personality can be defined by five dimensions known as Big Five personality traits. In general, study of personality considered as a psychology research based on the survey or questionnaire. But this limits the research data to less number of persons. Hence there is a need of something through which we can increase the number of people involved in survey and to make the process automated.

Data from Online Social Networking Sites provides a solution to this problem. The rapid growth in social media increased people perceptions towards it. It went from niche activity to vary widely and heavily used. It has emerged as one of the most ubiquitous means of communication today. It allows individual to find like-minded ones, whether it be for romantic or social purpose. It is also being used to maintain existing social connections. Tidwell and Walther observed that online interactions generated more self-disclosures and fostered deeper personal questions than did face-to-face conversations. Now-a-days people analyze person's social profile before considering as business partner or before dating. Researchers have shown how useful social networking is among old adults, what can we learn from Facebook activity and how often it is used by famous personality.

Ordinarily, by making the record on long range interpersonal communication site like Facebook individuals gives the privilege to gather their data; in light of this information, Facebook explores group attempt to screen the user conduct. Numerous

highlights and traits of these long-range informal communication locales are valuable for personality assessment. Specialists have demonstrated this and could evaluate the personality traits by the utilization of online networking, for example, Facebook and Twitter. In light of social behavior toward companions and in view of auxiliary data like various companions, bunches joined and likes and so forth can be utilized to effectively anticipate a portion of the personality traits. In this paper, we will endeavor to cover the investigates done to foresee the identity qualities utilizing online networking information, the calculation utilized, impediment and results.

Personality traits play an important role in identification of an individual and assessing a role or a responsibility of an individual. There are five types of personality traits namely neuroticism, extraversion, openness, agreeableness and conscientiousness.

There are multiple methods to understand personality traits but these methods are time consuming and so there comes a need of a quick way or framework that can be executed easily and the one that accepts natural habits and instant responses of individual. Social media is one of the most easily accessible ways to understand natural behavior of an individual, understand user's likes and dislikes and so we can link information extracted from social media to understand personality traits of social media users.

It has been found that lot of work has been done in the past to bridge the gap between social media and personality traits by using the information people reveal in their online profiles. It has been proven that social media can be used to predict personality traits. Amongst all social media sites, face book profiles are reflective of their actual personalities. Text analysis tools are used in the past to aggregate and quantify the data available on social media.

This paper is presented in three sections where Section 1 covers the related work done in the area of identifying personality traits including classic methods and making use of information available on social media network.

Section 2 explains the approach of the study and also explores the use of text analytics in social media to depict personality trait of the users.

Section 3 proposes a framework that can be implemented to identify the personality traits of social media users by accessing their likes, comments and shares on social media using text analytics.

II. RELATED WORK

Personality of an individual can be identified using big five personality traits also known as the five factor model. The five factor model is a well proven model based on five traits: openness to experience, conscientiousness, extraversion, agreeableness and neuroticism. The model came into existence after a wide research on personality and is well accepted worldwide.

Table 1: Characteristics of Big Five Personality Traits

Openness to experience	Appreciation of art, emotion, adventure, unusual ideas, curiosity, variety of experience, imaginative, independent, intellect
Consciousness	Self-discipline, dutifully, aim of achievement, planned, organized, dutifully, Idealism
Extraversion	Gregariousness, assertiveness or leadership, social confidence, Orderliness, Industriousness, Self-discipline, Energy, positive emotions, assertiveness, sociability, tendency to seek stimulation, talkativeness
Agreeableness	Modesty, trust, Empathy, Altruism, Compassionate ,cooperative, suspicious, antagonistic ,helping nature
Neuroticism	Ways to experience unpleasant emotions: anger, anxiety, depression, vulnerability, compulsiveness, Rumination

DETAILS ABOUT PERSONALITY TRAITS

Extraversion: It relates with person's tendency to get involved with external world. People high on extraversion tend to be more outgoing, friendly and socially active. Those with low score are likely to be solitary and reserved.

Agreeableness: It is a measure of maintaining positive social relationship. People high on Agreeableness tend to be cooperative, friendly, compassionate and adaptive. Low scorer are highly disagreeable people and are suspicious, distant and uncooperative the place.

Conscientiousness: highly conscientiousness people are like achiever; they are always good in discipline, responsible and prefer proper planning ahead. High Conscientiousness suggests a strong ability to regulate and control behavior.

Neuroticism: This is measure of emotional stability. Highly neurotic people are more prone to negative emotions like anxiety, anger, nervousness, stress and depression, more likely to be frustrated in day to day life. Low scorers are calm and collected, emotionally stabled and balanced.

Openness to experience: Relates to person's curiosity, interest in new experience/ideas, imagination. People with high score on this trait appreciate art, adventure and new ideas whereas ones having low score tend to be conservative, conventional.

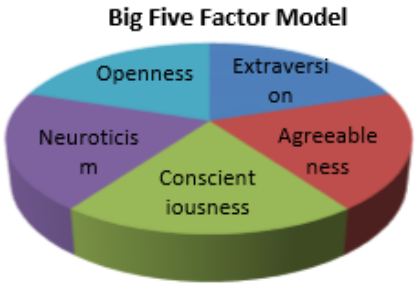

Figure 1: Five attributes of Big Five Factor Model

III. LITERATURE SURVEY

J. Golbeck et al. stated in their research that they were the first ones to look at the relationship between personality traits and social profile statistics. They created a Twitter application through which they undertook recent 2000 tweets of fifty subjects. The subjects were presented with the 45 question version of the Big Five Personality Inventory. The collected tweets corpus was processed with the help of two tools, first LIWC (Linguistic Inquiry and Word Count) from which they were able to extract a total of 79 features. Second, MRC Database which yielded 14 language features. They also performed a word by word sentiment analysis with the help of General Inquirer dataset. The authors were able to predict scores from 11% to 18% of their actual values.

C. Ross et. al. have used 28-item questionnaire related to Facebook uses and was able to relate it with personality traits of the individual.

T. Ryan et. al. have used 124 questions including Facebook usage questionnaire, the big five inventory consist of 44-items that yield Big Five personality trait scores. In this research it is proved that extraverted people are more likely to use Facebook than introverted people. Facebook users have higher level of narcissism. Individuals higher on neuroticism prefer using the wall.

Both the research and uses Facebook usage questionnaire. Some results are harmonious like people high on neuroticism use asynchronous communication and prefer using wall. However, they contradict with each other on some aspects. For example says communicative features of Facebook and it relation with Extraversion is positively correlated however is contradictory on this. This contradiction might be due to different input data. [39] is having homogenous data among university student with 15 male and 82 female while has conducted the research on 1635 self-selected internet users between 18 to 44 years old. This is one of the simplest approaches towards personality prediction. However, the questionnaires used in many big five personality studies are typically lengthy. Efforts have been made to develop brief scales in psychology. In this context Gosling et al. introduced Ten Item Personality Inventory (TIPI) that includes ten questions to determine the Big-Five personality traits. Many study like use this TIPI to measure self-perceived personality. It is more straightforward to ask a person how extraverted he is than to ask him whether he enjoys the company of others attends parties frequently, is talkative, outgoing, gregarious, and enthusiastic. Asking multiple questions for one trait is reduced by asking one question to avoid redundancy, boredom to give the answer and to reduce the time, so that more people can participate in survey.

Following are some limitations associated with personality questionnaire approach which may result into inaccurate or suboptimal results:

• People may fake quality for few minutes.

• Time consuming.

• Expensive as professionals charge per candidates.

• Accuracy as they are being judged by the person, so at times qualification and talent of person matters here. Besides, humans have natural tendency to prejudge so it is prone to human errors.

Traditional Methods used to identify personality traits – After the theory of Big five personality was put forth, several classic methods were used to understand personality of an individual.

These methods were implemented in form of data, ratings, self-reports, questionnaire, data from experimental settings. In the past personality traits were identified by means of selecting a random sample and conducting a survey or gathering information in form of a questionnaire. Using the Samejima's model, traits were estimated and the same was used to discriminate the individuals.

In another research, card game was used as method instead of questionnaire. Card game was played between sixty software practitioners .The outcome of the game was used to identify the personality trait of the individual with the help of MBTI scale. The same practitioners were also tested in the industry environment. Most of the people were found to be extroverted in the industry environment as well as outside the industry environment.

IV. APPROACH

Data Mining Techniques:

With growing use of internet, lot of information is being shared worldwide. There is a need to analyze this information by using appropriate techniques to recognize patterns in available information. This implies that use of Data mining techniques is essential in identifying personality trait through social media in the world of internet.

Data mining techniques have been applied to some parts of social media. Major information shared through social media network is in form of text. This text shared by millions of users can be categorized using several demo graphs like group of users belonging to a particular geographical location or belonging to a particular gender or users that fall in a particular age-group. After categorization, we can analyze this textual information using data mining techniques. One of the widely used techniques in data mining is text analytics that is best suited for the area of social media. Here, the focus is basically on retrieving some important information from the available text. It generally includes categorization of the text based on the requirement, extracting the concept hidden within that text.

The Text mining approach is a wide area into itself. A lot of classification algorithms and decision tree algorithms can be applied in this area. Miner, Gary in his book suggests strongly the area of document classification and concept extraction in web mining.

Sentiment analysis is another area in text mining where one can find the emotions of users based on the text they share. Jim Sterne in his paper states that the computer has the ability to perform the sentiment analysis on the text using the tools and techniques. So, by using the phrases included the text, we can identify whether it is used in positive aspect or negative aspect. The mood or emotion of the individual can be thus understood from the text.

Text Analytics in Social Media:

Data available from the social media can be in the form of text, images, blog or web page. Here, we are restricting our research towards the text and therefore will be using text analytics as a tool. Text analytics is useful in deriving better quality inferences from the collected text. Better quality in the considered scenario means combination of relevance, interest and novelty.

Text analytics is the process of accepting input text, structuring the text, deriving patterns within that text and interpreting the output. Here, the challenge is to apply proper text analytical method for data analysis so as to properly interpret the text used in comments. In general, the text analytics is used to perform sentiment analysis on social media data.

By observing the document, the expressions used in the document and also by observing the words used in the document, the associated sentiments can be predicted. The approaches that can be used for the sentiment analysis can be Natural Language Processing (NLP) & pattern-based approach, Machine learning algorithm, Hybrid Classification etc.

The classifiers used for the classification of the sentiments are General Inquirer Based Classifier (GIBC), Rule Based Classifier (RBC), Statistics Based Classifier (SBC) and Induction Rule Based Classifier (IRBC).

Rule Based Classifier is consists of if-then relation. LHS of the rule will be the condition and RHS of the rule will be the result. If the condition is satisfied by the data during analysis, then the data can be considered into the category specified on the RHS of the rule.

For analyzing the sentiments associated with the text used in the comments by individual, rule based classifier can be used. The classifier consists of the rules where the condition will the combination of the words/phrases included in the comments while the result will be the associated sentiment.

V. PROPOSED FRAMEWORK

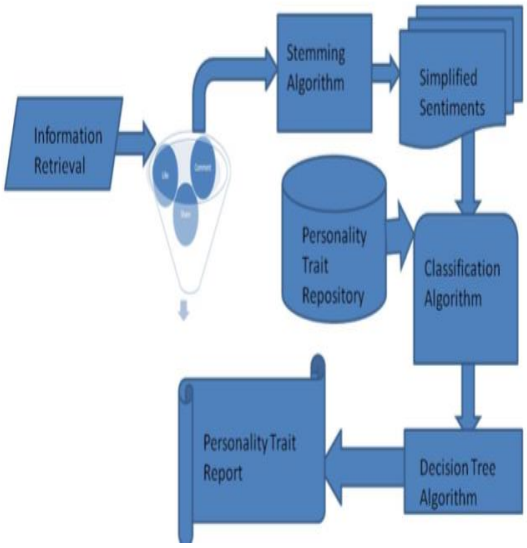

Figure 2: Proposed Framework to Identify Personality Trait of Social media users: A data mining approach

Step 1: Data Filtering

Data collected from the social media will be the text from comments posted by the user. The filtering of the text requires some phrase and pattern based techniques or term based techniques.

Here, the phrase based technique is preferred because phrases carry more semantic information than terms and hence better performance can be expected. The main aim for filtering data is to remove the redundant or irrelevant data. As a result, we will get clean data which can be processed more effectively.

First of all, the probable phrases and their synonyms that can occur in the comments are listed. This list helped in extracting those phrases from the text. Also, the dictionary including list of words like ‘a’, ‘an’, ‘the’, ’you’, ’of’, ’over’ etc. is made to avoid useless text from getting processed.

Step 2: Data Stemming

Data stemming uses the extracted phrases after data filtering. Stemming is the process for reducing the words to their stem or root form. In this, the set of words that can be treated as equivalent are identified and these multiple occurrences are replaced with their root form.

There are many stemming algorithms that can be used to serve the purpose like Lookup algorithms, the production technique, Suffix-stripping algorithms, stochastic algorithms, Porter stemming algorithm, matching algorithms etc.

Porter stemming algorithm is one of the most popular algorithm that is used for data stemming. This algorithm removes the suffixes from the words that have been added to the right-hand end of root form.

Step 3: Simplified Sentiments

After data stemming is done, the input will be provided for simplifying the sentiments. The sentiments which are associated with the text used in comment may be positive, negative or neutral. The input here is the stem or root form of the words or phrases used in the comments. So, it is easier to identify the corresponding sentiments.

Step 4: Personality trait repository

Personality trait repository is used to associate the Big five personality traits with the corresponding attributes. The attributes considered here are listed in Table 1. Each attribute included in the repository is again linked with the synonymous words. The information retrieved is the text in comments.

The text is composed of phrases, certain adjectives, smiley and also some punctuation. These phrases and adjectives will be the input to the repository where association between phrases or adjectives and synonymous words will take place.

Step 5: Classification Algorithm

There are different existing algorithms that can be used for text analytics like Classification Algorithm, Association Algorithm, and Clustering Algorithm. All these algorithms have their own advantages as well as limitations.

Classification algorithm deals with assigning a keyword to document based on a defined keyword set. It requires collection of records where each record has unique record id and fields corresponding to attributes. Methods used for text classification can be Decision trees, Pattern classifiers, SVM classifiers, Neural Network classifiers, Generative classifiers etc.

Here, the study involves use of Pattern/rule based classifiers for classification algorithm. The pattern/rule based classifier determines word patterns which are most likely related to the different classes. Researchers have constructed a set of rules where each rule is associated with a keyword.

A person cannot be strictly categorized to belong to one of the personality trait. However, a person can have a combination of the characteristics that belong to the five personality traits as explained in table 1. The percentage of those characteristics will vary based on the responses of the user for the post. The personality trait with highest weight-age among the five personality traits can be treated as his/her personality trait.

Step 6: Decision tree algorithm

Decision trees are found to be powerful and popular tools for classification and prediction. Decision trees represent rules which can be easily understood by anyone and at the same time, it can be used in a database system. This algorithm requires attribute-value description and pre-defined classes.

The properties of the attributes are collected and provided as input to decision tree algorithm. Also, the pre-defined classes from the classification algorithm are provided to the decision tree algorithm. The rules defined here are used to derive results in terms of personality traits. This can further be used to create personality trait report.

VI. CONCLUSION

Predicting personalities using likes, comments and shares is surely a real life problem due to its vast applications in diverse fields and must be recognized as a significant field of study under natural language processing and must be harnessed with the predictive potential of machine learning. A lot of work is still to be done which can only be accomplished by overcoming the constraints put forward by language use and intent of users based on their own choices. In this paper, we intend to put forward the need of research communities to come forward and gather enough resources to make machine learning a feasible method for prediction on both macro and micro levels.

REFERENCES

[1] M. Back, J. Stopfer, S. Vazire, S. Gaddis, S. Schmukle, B. Egloff, and S. Gosling. Facebook Profiles Reflect Actual Personality, Not Self-Idealization. Psychological Science, 21(3):372, 2010.

[2] Golbeck, J., Robles, C., & Turner, K. (2011, May). Predicting personality with social media. In CHI'11 Extended Abstracts on Human Factors in Computing Systems (pp. 253-262). ACM

[3] H.V. Zhao, W.S. Lin and K.J.R Liu, "Behavior modeling and forensics for multimedia social networks" Proc. Signal Processing Magazine, IEEE (Volume:26 , Issue: 1), pp. 118-139, 2009.

[4] Sumner, A. Byers, R. Boochever, & G.J. Park, "Predicting dark triad personality traits from Twitter usage and a linguistic analysis of tweets", in 11th international conference on machine learning and applications, pp. 386–393, 2012

[5] Asur, Sitaram, and Bernardo A. Huberman. "Predicting the future with social media." Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on Vol 1 IEEE, 2001

[6] Cambria, Erik, et al. "New avenues in opinion mining and sentiment analysis. "IEEE Intelligent Systems 28.2(2013): 15-21

[7] Golbeck, Jennifer, Cristina Robles, and Karen Turner. "Predicting personality with social media." CHI'11 extended abstracts on human factors in computing systems. ACM, 2011. [8]. Barrick, Murray R., and Michael K. Mount. "The Big Five personality dimensions

and job performance: A meta-analysis." (1991).

[9] Judge, Timothy A., et al. "The big five personality traits, general mental ability, and career success across the life span." Personnel psychology 52.3 (1999): 621-652.