Personality Prediction using Twitter Data

Sanjit Kumar R[1], Shrivatson R G[2], Rishi Priyan S[3], Padmavathy T[4]

[1,2,3]*Undergraduate, Dept. of Computer Science and Engineering Engineering, Sri Venkateswara College of Engineering, Tamilnadu, India*

[4]*Assistant Professor, Dept. of Computer Science and Engineering Engineering, Sri Venkateswara College of Engineering, Tamilnadu, India.*

Abstract - *Social media has grabbed the attention of the current generation people and became very much accessible on the internet. The activity performed by social media users provides a better platform for researchers to study and learn about online behaviors, preferences, and personality traits. We are using Twitter as a leading social media platform for the analysis since the number of users is intense. To analyze the person's nature, the user data can be extracted and perform further analysis based on the text, status profiles, or preferences provided to other users on Twitter. To understand the different forms of texts, the comment added by the user, tweet, social behavior, and the usage of language habits on Twitter is taken as the data for analysis. This research can be taken further by using Machine Learning algorithms to process the data and the user's actual prediction personality.*

Key Words: Social Media, Personality Traits, Machine Learning.

1. INTRODUCTION

Social network platforms play a significant role by seizing the attention of persons active on the Internet. A group of researchers has found that many persons are active on the social media platform so that they can discover the different personalities of various people in their daily lives. This information can be gathered employing comments, tweets, mutual connections, and any form of text messages. By using these kinds of processed data can denote the user's online behaviour on other profiles. Initially, our study focuses on social behaviour and language habits on twitter. Secondly, we choose the most useful features for each personality dimension and predict the user's character. We proposed a systematic system based upon two categories. One is Social network analysis(SNA) and two classes of Linguistic Inquiry and word count(LIWC) and Structured Programming for Linguistic Cue Extraction(SPLICE) based on the dataset we used. Based on these categories, we can check some of the similarities shown by processing the data. The correlations between the personality traits and feature sets are nearly the same. Several basic standards predict easier to predict. Using several sets of categories such as network size, the density of tweets, profession, and the number of connections, we came into a conclusion to process the data to information. We have explored the predictive nature of the features set by predictions using personality traits. We have scrutinized a lot of information, which is undesirable for the further process so that the accuracy and the efficiency of the forecast can be higher. Finally, we have used a parallel Machine Learning algorithm to proceed with the implementation model to traverse the extent to which we can predict the personality traits from twitter. We have studied and chosen the best algorithm for a higher prediction. In our research, we exploit data collected utilizing twitter.

Our research, have some contributions towards to predict an efficient data: first, to study about the connections between the users and their interactions in a social network; second, to elucidate the social network features between different categories using XGBoost machine learning approach; third, the relationship between the categories; fourth, proving that SNA provides higher efficiency than other elements[1].

2. Literature Survey

2.1 The development and psychometric properties

Pennebaker et al. [2] explained the work of personality extraction from the text. They have collected the data through a manual process such as dairy writings, college assignments, and psychology manuscripts to study the personality of a person. Their outcomes are little satisfied for extravert person and not for introverts. Usually, introverts don't show much interest in these practices. So, this fails in the broader area.

2.2 Lexical predictors of personality type

Argamon et al. [3] worked on extraversion and neuroticism using linguistic features using some categories of traits. They are appraisal, function words, expressions, modal verbs, etc. In this work also, failure is caused due to some lesser outcomes. It lags in the result of extraversion and mainly used for functional lexical features.

2.3 Personality, coping, and coping effectiveness in an adult sample

Oberlander and Nowson [4] classified the personality traits using the style of language habits from bloggers using a Machine learning algorithm called Naïve Bayes to predict it. This has been applied using different sets of n-gram as features, and it is feeble in the copying of efforts on well-being personality[6].

2.4 Psychological predispositions and network structure: The relationship between individual predispositions, structural holes and network closure

Kalish and Robins [5][7] examined the personalities of individuals by working on their immediate network. They both were working on the egoist character of the environment. By focusing on this, they were successfully collected the datasets of more person's character. Their analysis concluded that psychological predispositions could be explained by the dataset even though the text or language-based data has received much attention and more efficiency through other research.

3. PROPOSED WORK

In this research, we have obtained the dataset from Twitter social media platform. These data are preprocessed and passed into Feature Extraction and then Feature Selection. And finally, the extracted features are trained by the XGboost algorithm, and the resultant classifies the personality with the accuracy, as mentioned in the architecture diagram.

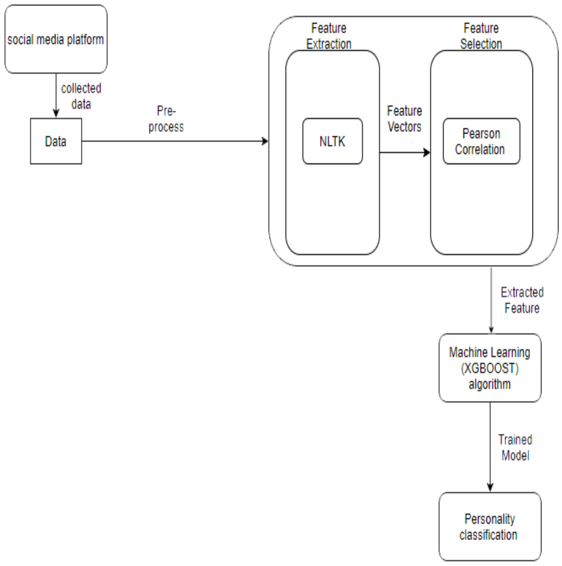

Fig -1: Proposed Architecture Diagram

3.1 Dataset Description

Data set are being collected from the twitter API . It let us to make use of features of twitter without the help of website interface. Twitter API can be used for posting tweets or sending message in an automated way. API is also known as a set of documentation of URLs. Through this, we have several data to predict the people's personality.

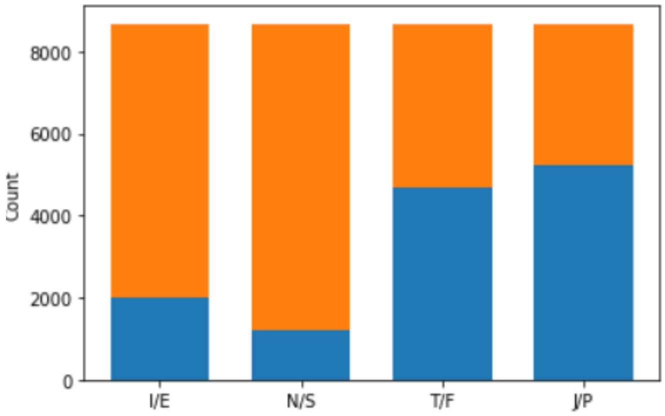

Chart -1: Distribution across types Indicators

3.2 Data Preprocessing

The dataset gathered from my Personality was pre-processed before it progressed to the feature selection and training stage. To pre-process the dataset, we used OpenNLP. First, we utilized tokenization to divide the last word of every sentence, including punctuation and aggregation of identical terms. Next, we removed URLs, symbols, names, spaces, and lower cases.

3.3 Feature Extraction

The people behaviour on social media network is correlatively affected due to the behaviour and the presence of other people on the platform. This can create an influence on the development of new data or actions through the groups. Many applications have been developed to understand how these actions arise and affect the people. In our research we classify the dataset into two groups, the first one is text feature extraction which is employed to analyse the people's language on Twitter it also includes two features namely topic and expression counts[8]. The LIWC(Linguistic Inquiry and Word Count) is potentially used to understand the psychology of the people, this tool is vast spread for psychological studies. LIWC2015 is mainly developed to analyse the files in many languages in a timely period and efficiently[9]. SPLICE(Structured Programming for Linguistic Cue Extraction) it is a analysis tool for linguistic which is under testing stage. It can be used for personality prediction when it is completely developed. Feature extraction is done by using NLTK(Natural Language ToolKit).NLTK is an nlp library that comprises packages to do understand human communication to the machine. Package includes stemming, lemmatization, tokenization, POS tagging, extraction[10].

3.4 Feature Selection

Usually, there are a couple of main reasons how feature selection is essential for developing a model[11]. It lessens the high dimensionality of the dataset by eliminating the characteristics not necessary for training, strengthening the model's generalization, also reducing the training time. Then the model attains a sound knowledge of the features and their relations to the response characteristics. Added with it, it also increases the accuracy of the training algorithms and decreases the processing specifications[12]. We have used pearson correlation method for evaluating the linear dependency between two quantitative or continuous variable. It is also referred as Pearson R statistical test. Here the value returned is between -1 to +1.

- If the value is -1, it denotes a well built negative relationship between the variables.
- If the value is 0, it denotes no relationship between the variables.
- If the value is 1, it denotes a well built positive relationship between variables.

$$r = \frac{n(\sum az) - (\sum a)(\sum z)}{\sqrt{[n\sum a^2 - (\sum a^2)][n\sum z^2 - (\sum z)^2]}} \quad (1)$$

In eqn.(1) r is Pearson Coefficient, n is the number of pairs of variable, $\sum az$ is sum of product of the paired variables, $\sum a$ is sum of the a variable, $\sum z$ is sum of the z variable, $\sum a^2$ is sum of the squared a variable, $\sum z^2$ is sum of the squared z variable.

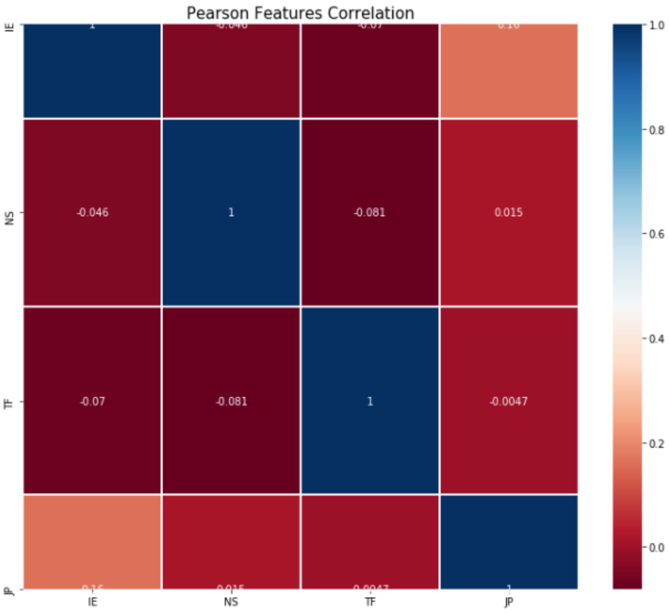

Chart -2: Applying Pearson Correlation to the Data.

3.5 XGBoost Algorithm

XGBoost algorithm is making use of the thought of gradient tree boosting. This algorithm was developed to enhance the speed and performance higher.

$$\hat{y}_i = \sum_{t=1}^{m} f_t(x_i) \quad (2)$$

In order to calculate the gain we have to minimize the loss function. Loss can be calculated as follows,

$$Loss = \sum_{i=1}^{n} L(y_i f_{m-1}(x_i) + h_m(x_i)) + \sum_{j=1}^{T} \Omega\left(h_m(x_j)\right) \quad (3)$$

Irrespective of the loss function, the L functions which is the expansion of the Taylor series is employed. Then the regularization term Ω can be written as,

$$\Omega_m(h_{mT}) = \in T + \frac{1}{2}\lambda \sum_{j=1}^{T} b_{jm}^2 \quad (4)$$

The resultant gain after splitting the leaf into two leaves is calculated as below,

Gain = Loss Before Split-Loss After split (5)

We have to apply the XGboost algorithm with irregular individuals of data with those collected data. In which we use bagging to produce irregular individuals from the dataset and equip each sample independently and given with an average value for each tree. Then in boosting the average values of each tree is get replaced with an over-weighted value. Here we have a complete explanation of bagging and boosting.

Bagging is a simple and more efficient ensemble method. Bagging is the bootstrap procedure application which deals with a huge variant ML algorithm commonly decision trees[14].

- Take R random samples of a% of the samples and b% of the features from the data set.
- Place the tree on each of R random samples.
- Predict with each R random sample.
- Final prediction can be known from the average predictions.

Boosting is defined as a group of algorithms that make weak learners into stronger learners by utilizing weighted averages. Boosting works in the principle of teamwork.

- Place the tree into the data.
- Get the residuals.
- Place the tree to the residuals.
- Move to second step for R boosting rounds.

- The weighted amount of the sequential predictors is taken as a final prediction.

3.6 Personality Classification

To establish the personality classification of a user, we use Myers- Briggs. The MBTI is the most commonly used for character categorization, which has been used in several departments to analyze the individual's personality [15]-[17]. MBTI classifies the character based on four-category, namely:

- Introversion (I) or Extraversion (E)
- Intuition (N) or Sensing (S)
- Feeling (F) or Thinking (T)
- Perceiving (P) or Judging (J)

Type	Expansion
ISTJ	(Introverted, Sensing, Thinking, Judging)
ISFJ	(Introverted, Sensing, Feeling, Judging)
INFJ	(Introverted, Intuitive ,Feeling, Judging)
INTJ	(Introverted, Intuitive ,Thinking , Judging)
ISTP	(Introverted, Sensing, Thinking, Perceiving)
ISFP	(Introverted, Sensing, Feeling, Perceiving)
INFP	(Introverted, Intuitive, Feeling, Perceiving)
INTP	(Introverted, Sensing, Thinking, Perceiving)
ESTP	(Extraverted, Sensing, Thinking, Perceiving)
ESFP	(Extraverted, Sensing, Feeling, Perceiving)
ENFP	(Extraverted, Intuitive, Feeling, Perceiving)
ENTP	(Extraverted, Intuitive, Thinking, Perceiving)
ESTJ	(Extraverted, Sensing, Thinking, Judging)
ESFJ	(Extraverted, Sensing, Feeling, Judging)
ENFJ	(Extraverted, Intuitive, Feeling, Judging)
ENTJ	(Extraverted, Intuitive. Thinking, Judging)

Table-1: Types of Personality Indicator

- Introversion (I) or Extraversion (E)
  People and Things in the external world is mainly focused by an Extraverts, whereas an introverts look into inner views and ideas.

- Intuition (N) or Sensing (S)
  Sensing refers to recognize objects through vision, noise, flavor,tap, and smell, while intuition refers to presiding type such as looking into the previous experience and also concentrate more on their thinking.

- Feeling (F) or Thinking (T)
  The thinking and feeling come into the category called judgment where we can judge people. Thinking uses sense to decide the society, where feelings lead us to outlook objects base on what sentiment they evoke.

- Perceiving (P) or Judging (J)
  Every person evaluate and discern, but person who are judging presiding are known to be well ordered and outcome-driven, where discern presiding are flexible and multitasking personality.

4. EXPREMENTAL ANALYSIS

The project determines a person's personality by using posts that are available on their respective Twitter accounts, which can be used in the recruitment process in companies. Our result illustrates a greater accuracy of 81.23% of the person's personality through the result of our model, which is trained using the Extreme gradient boosting algorithm by classifying using the Myers Briggs Type Indicator.

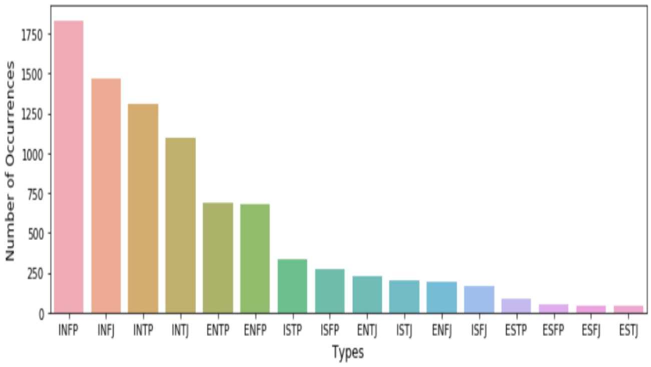

Chart -3: Personality Traits

5. CONCLUSIONS

In this research, we have provided a framework of insights on every user's social network analysis and personality prediction. Our survey on this regular operation of study is made successfully by using the user's perspective on the relationship, tweets, and comments in the social network. To predict a user's nature or character, we have supervised a comparative study of best behavioral standards for the Twitter usage of the same set of features to record the ways the user's socialized, communicate and connect. To determine the different personalities of a person, we have used a broad set of datasets for an efficient outcome. Our outcomes show that a more significant number of insights can be gained from analyzing the social and linguistic network of personality.

6. FUTUREWORK

By inspecting the personalities from Twitter, profile statuses may also recommend some similar conditions to predict an individual's profile like TV shows, music, or sports. The items could be endorsed to individuals based on mutual connections also. We can also recommend similar items to the bilateral users who are all in the same nature. This type of recommendation can be done by a collaborative filtering method, and it could apply to users who share similar thoughts. Developing and performing these operations of approaches in our personality recommendation system could be our research's future work.

REFERENCES

[1] J. Staiano, B. Lepri, N. Aharony, F. Pianesi, N. Sebe, and A. Pentland, "Friends don't lie: Inferring personality traits from social network structure," in Proc. ACM Conf. Ubiquitous Comput., 2012, pp. 321–330.
[2] J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn, "The development and psychometric properties of LIWC2015," Tech. Rep., 2015.
[3] T. P. Michalak, T. Rahwan, and M. Wooldridge, "Strategic social network analysis," in Proc. AAAI, 2017, pp. 4841–4845.
[4] S. Argamon, S. Dhawle, M. Koppel, and J. Pennebaker, "Lexical predictors of personality type," Tech. Rep., 2005.
[5] Y. Kalish and G. Robins, "Psychological predispositions and network structure: The relationship between individual predispositions, structural holes and network closure," Social Netw., vol. 28, no. 1, pp. 56–84, 2006.
[6] R. R. McCrae and P. T. Costa, "Personality, coping, and coping effectiveness in an adult sample," J. Pers., vol. 54, no. 2, pp. 385–404, 1986.
[7] P. T. Costa and R. R. McCrae, "Normal personality assessment in clinical practice: The NEO personality inventory," Psychol. Assessment, vol. 4, no. 1, pp. 5–13, 1992.
[8] S. Adali and J. Golbeck, "Predicting personality with social behavior," in Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM), Aug. 2012, pp. 302–309.
[9] D. J. Brass, "Being in the right place: A structural analysis of individual influence in an organization," Admin. Sci. Quart., vol. 29, no. 4, pp. 518–539, 1984.
[10] R. Wald, T. M. Khoshgoftaar, A. Napolitano, and C. Sumner, "Using Twitter content to predict psychopathy," in Proc. 11th Int. Conf. Mach. Learn. Appl. (ICMLA), vol. 2, Dec. 2012, pp. 394–401.
[11] M. A. Hall, Correlation-Based Feature Selection for Machine Learning. 1999.
[12] G. Farnadi et al., "How are you doing?: Emotions and personality in Facebook," in Proc. EMPIRE (2nd Workshop Emotions Personality Personalized Services), Workshop UMAP (22nd Conf. User Modeling, Adaptation Personalization), 2014, pp. 45–56.
[13] S. Bai, R. Gao, and T. Zhu, "Determining personality traits from Renren status usage behavior," in Computational Visual Media. Springer, 2012, pp. 226–233.
[14] H. Wei et al., "Beyond the words: Predicting user personality from heterogeneous information," in Proc. 10th ACM Int. Conf. Web Search Data Mining, 2017, pp. 305–314.
[15] Y. Amichai-Hamburger and G. Vinitzky, "Social network use and personality," Comput. Hum. Behav., vol. 26, no. 6, pp. 1289–1295, 2010.
[16] M. Selfhout, W. Burk, S. Branje, J. Denissen, M. Van Aken, and W. Meeus, "Emerging late adolescent friendship networks and big five personality traits: A social network approach," J. Pers., vol. 78, no. 2, pp. 509–538, 2010.
[17] K. J. Klein, B.-C. Lim, J. L. Saltz, and D. M. Mayer, "How do they get there? An examination of the antecedents of centrality in team networks," Acad. Manage. J., vol. 47, no. 6, pp. 952–963, 2004.

BIOGRAPHIES

R. Sanjit Kumar,
Student,
Department of Computer Science,
Sri Venkateswara College of Engineering,
Chennai-602117

R.G. Shrivatson
Student,
Department of Computer Science,
Sri Venkateswara College of Engineering,
Chennai-602117

S. Rishi Priyan,
Student,
Department of Computer Science,
Sri Venkateswara College of Engineering,
Chennai-602117

T. Padmavathy,
Assistant Professor,
Department of Computer Science,
Sri Venkateswara College of Engineering,
Chennai-602117