International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 05 | May-2018 www.irjet.net p-ISSN: 2395-0072

A Comparative Study of Different Classifiers for Myers-Brigg Personality Prediction Model

Shristi Chaudhary[1], Ritu Singh[2], Syed Tausif Hasan[3], Ms. Inderpreet Kaur[4]

[1,2,3,4] *Department of Computer Science and Engineering, Galgotias College of Engineering and Technology,Greater Noida, India*

Abstract—*Personality can be understood as specific features of an individual which determines its preferences over things. Personality has been shown to be relevant to many types of interactions such as in predicting recommendations e.g. movie preferences, social interactions, music preferences, criminal activities and correlation between personality and job performance. Predicting personality from social media become one of the most trending things among researchers and various commercial organizations as to help individual improve their experience over computerized user interfaces and help others to study the various personality preferences. Thus, many algorithms have been performed to predict personality from social media. In this paper, we compared the performance of several classifiers such as Naïve Bayes, Random Forest etc. in predicting Kaggle Users personality. Based on the user profile and comments, the user data of Kagglers were extracted from Kaggle Repository, analyzed, and then classified in the automatic personality prediction. The user data is extracted and mapped on the Mayer's – Brigg personality Model. All sixteen co-ordinates of the MB-Model was considered in this study. A 5-fold cross validation was used to evaluate the classifiers. Several parameters that were observed in the performance of the classifiers are classification accuracy, F-measure, Logarithm-Log function. Experimental evaluation demonstrated that Logistic Regression algorithm is the best classifier in terms of the accuracy and F-measure.*

Key words: Personality Prediction, Natural Language Processing, Machine Learning, Myers-Briggs Model, Parameter Tuning

1. INTRODUCTION

According to [1], personality is defined as a set of attributes that describes an individual's uniqueness of behavior, temperament, emotion, and mental. In simple terms, personality represents the mixture of features and qualities that built an individual's distinctive character. There are many different personality models used to characterize personality such as the Big Five model (Five-factor model) [2], the Myers-Briggs Type Indicator (MBTI) [1], and the Theory of Personality Types Carl Jung [4]. In this study, the Myers-Brigg Type Indicator is selected for it is one of the least explored personality model [5]. The comparative study of the classifiers on the Extroversion personality on Big Five model [6] is explored by Nor Rahayu Ngatirin, Zurinahni Zainol, Tan Lee Chee Yoong.

Social platforms are places where people spend considerable amount of time in sharing personal life happening and surrounding happenings, communicate, and interact with others in exchange of knowledge and entertainment. Due to this, social platforms have become one unique source of big data where the information can be used to improve the living. According to [6], the uniqueness of social media data calls for novel data mining techniques that can effectively handle user-generated content with rich social relations. The research and development of these techniques are known as social media mining, an emerging discipline of data mining. Social media mining is the process of representing, analyzing, and extracting actionable patterns from social media data [7]. There are many different mining techniques were developed to mine these semi structured data from social media including Naïve Bayes, classification trees, and association rules. Various mining purposes have been performed in the social media sites for the purpose of extracting useful information on the behavior of users. The aim of this paper is to discuss the techniques that have been performed to predict user's personality under the Myers-Brigg personality model and compare the performance of the classifiers in order to get the most significant classifier for predicting user's personality. In the following section is the background study and theory of the Myers-Brigg Type Indicator.

2. BACKGROUND STUDY

Personality is described as a fairly fixed feature of an individual which indicates individual's preferences and may influence his/her decision making. It distinguishes an individual from others in characteristic patterns of thinking, feeling, and behaving. Efforts were put in generating a descriptive personality model or taxonomy in which personality can be understood in a simpler way [6]. Study on the personality have always been the topic of interest for psychologists and sociology, and one such experiment has been performed by the psychiatrist Carl Jung named as "Myers-Briggs type indicator". The Myers-Briggs Type Indicator (MBTI) is based on Carl Jung's theory of

psychological type. It indicates your personality preferences in four dimensions:

1. Where you focus your attention – Extraversion (E) or Introversion (I)

2. The way you take in information – Sensing (S) or Intuition (N)

3. How you make decisions – Thinking (T) or Feeling (F)

4. How you deal with the world – Judging (J) or Perceiving (P)

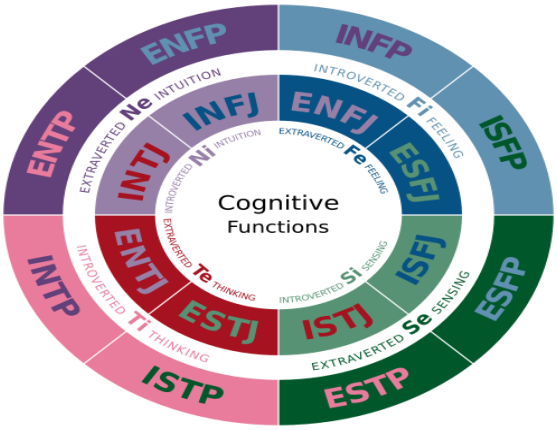

Fig-1: The 16 personality co-ordinates for an individual's personality [5]

The four letters that make up individual's personality type over the 16 personality co-ordinates.

3. RELATED WORK

Pennebaker and King (1999) examined stream-of consciousness (SOC) writings in terms of linguistic dimensions and personality trait. For their experiment, they utilized the Five Factor model, i.e. personality traits, as opposed to the MBTI personality types.

Also, here is the related work on the personality prediction system:

Table-1: Related work researches

Research	Social Sites	Mining Technique	Purpose
Alam et al. (2013) [20]	Facebook	SMO, BLR, and MNB	Personality Prediction
Lima et al. (2013) [20]	Twitter	Naïve Bayes	Personality Prediction
Celli and Polonio (2013) [21]	Facebook	Linguistic analysis of text	Personality and online interactions
Gou et al. (2014) [22]	Twitter	Lexicon-based approach	Personality prediction and sharing preference
Nie et al. (2014) [23]	Microblog	Linear semi supervised Regression	Personality prediction
Chen et al. (2015) [24]	Twitter	Linear regression	Personality prediction for advert. targeting
Nor Rahayu Ngatirin (2016)	Twitter	Bayes, Functions, Rules, Trees, Meta	Classifier comparison
Tommy Tandera (2017)	Facebook	SVM	Personality Prediction

4. METHODOLOGY

4.1 Dataset: Data was collected from the Kaggle repository of the user comments on the site with the labels of personality co-ordinates already assigned to them. The frequency of dataset is 8676 comments of different users with their unique author ID.

4.2 Data Visualization: Words per comment

Variance of words and the length of sentences were examined in this step to get the intuitive idea of the sentence structure for each personality co-ordinate.

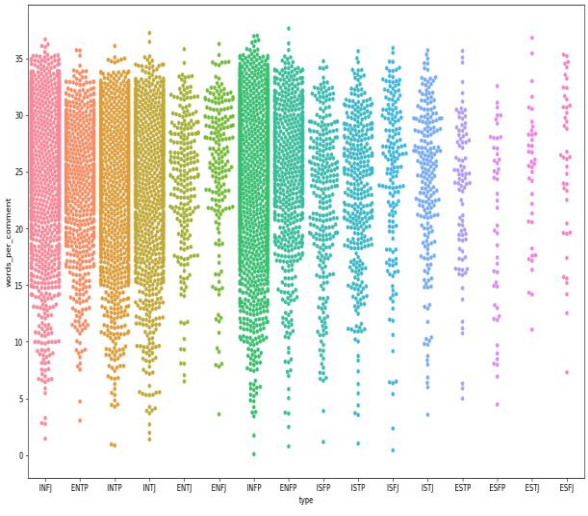

Fig -2: Words per Comment for each Personality type

a. Data Preprocessing: The data collected was crude e.g. https, .com, .xml annotations were contained.

b. Data Cleaning: Data was cleaned accordingly to avoid redundant results and Stop Words are removed.

c. Feature Extraction: In this process, Term-Frequency Over Inverse Document Frequency i.e. tf-idf feature extraction was taken into account, which reduced the probability of terms occurring in every document such as 'a', 'an', 'the'. The N-gram (unigram and bigram) are calculated to check the medium frequent words.

The formula for Estimating Bigram probability is given as,

- The Maximum Likelihood Estimate

$$P(wi \mid wi\text{-}1) = \frac{count(wi"1,wi)}{count(wi-1)}$$

$$P(wi \mid wi\text{-}1) = \frac{c(wi"1,wi)}{c(wi"1)}$$

4.3 Classification: In this phase, we performed different classification techniques to the data created in the previous phase in order to classify and predict the user's personality accordingly. A 5-fold cross validation was used to check the evaluation of the model. The classifiers along with their selected algorithms are given in table 2.

Table-2: Classifiers with selected algorithms

Classifiers	Algorithms
Bayes	Multinomial Naïve-Bayes
Functions	Logistic Regression, SVM
Tree	Random Forest

5. EXPERIMENTAL RESULTS AND DISCUSSION

To evaluate the performance of the classifiers, we have selected the algorithms given in table 2. we have divided over dataset in 5-fold cross validation and trained classifiers on them.
The table 3 depicts the scoring measures i.e. Accuracy and F-measure of the classifiers used.

Table-3: Accuracy and F-measure of the classifiers

Classifier	Accuracy	F-value
Extra Tree Classifier :	0.2971 (+/- 0.0192)	0.2971 (+/- 0.0192)
Naive_Bayes :	0.5589 (+/- 0.0097)	0.5589 (+/- 0.0097)
Logistic Regrssion :	0.6659 (+/- 0.0090)	0.6659 (+/- 0.0090)
SVM :	0.6544 (+/- 0.0124)	0.6544 (+/- 0.0124)

From the table 3, we can observe that the logistic regression is the classifier having the highest accuracy (with lowest standard deviation) 66.5% and F-measure .66. Whereas, Support vector machine have the significant accuracy of 65.4% and F-measure .65. Extra tree classifier is the worst performing classifier in our model.

We observed the F-measure of the selected algorithms. F-measure indicates the balance between the precision and the recall performance measures. It shows the preciseness of a classifier (how many instances it classifies correctly), as well as the robustness (it does not miss a significant number of instances).

As, we are considering Accuracy of the classifier as the main determining factor in selecting the classifier of our model, so we will train and test the model on logistic regression.

The other motive of the project was the optimization of the selected classifier i.e. parameter tuning.

Since, we are working on logistic regression, so we will tune the parameters accordingly which are Regularization parameter, C and N-gram model in the given case. The Table 4 depicts the improved accuracy of our proposed model.

Table-4: Result Comparison after Parameter Tuning

Logistic Regression	Original Accuracy	After Parameter Tuning
Accuracy (%)	66.59	67.75

From the table 4, it is observed that by adjusting optimized parameters for the classifiers, accuracy is increased by more than 1%.

6. CONCLUSION AND FUTURE WORK

The results of project shows that the machine learning algorithm i.e. Logistic Regression can improve the accuracy by tuning its parameters accordingly and the accuracy estimates are considerably good. It is possibly due to considerably small number of dataset used in this study.

However, the results are centered over the traditional machine learning algorithm and its performance can differ when performed by deep learning and other modern techniques. Also, our sample data don't come from all Kaggle user population, it comes from Kaggle users who write comments so, our conclusion can't be applied to all users who write over social sites, only to those who write comments.

Hence, for future study, we plan to collect and build more dataset. We also plan to use XGBoost algorithm 20, ther

architectures, and other processes to improve this prediction system.

REFERENCES

The Myers & Briggs Foundation. (2016). MBTI®Basics.[Online].Available:

[1] http://www.myersbriggs.org/my-mbtipersonality

[2] L. R. Goldberg, "The Structure of Phenotypic Personality Traits," American Psychologist, vol. 48, pp. 26-34, 1993.

[3] F. Mairesse, M. A. Walker, M. R. Mehl, and R. K. Moore, "Using Linguistic Cues for the Automation Recognition of Personality in Conversation and Text," Journal of Artificial

[4] Intelligence Research, vol. 30, pp. 457-500, 2007.

[5] C. G. Jung. (1921). Psychological Types. [Online].Available: http://ccmps.net/jung/types.pdf

[6] J. Golbeck, C. Robles, K Turner, "Predicting personality with social media," in Proc. of 2011 Annual Conference Extended Abstracts on Human Factors in Computing System. Vancouver, BC, Canada, 2011, pp. 253-262.

[7] [Nor Rahayu Ngatirin, Zurinahni Zainol, Tan Lee Chee Yoong, "A comparative study of different classifiers for automatic personality prediction" in 2016 6th IEEE International Conference on Control System, Computing and Engineering, 25–27 November 2016, Penang, Malaysia

[8] R. Zafarani, M. A. Abbasi, and H. Liu. (2014). Social Media Mining: An Introduction (Draft version), Cambridge University

[9] Press.[Online].Available: http://dmml.asu.edu/smm

[10] https://www.opp.com/en/tools/MBTI/MBTI-personality-Types, as dated on: 11.04.2018

[11] Jake Beech. (2016). Myers Brigg Type Indicator,[Online].available: https://commons.wikimedia.org/wiki/File:Cognitive_Functions.svg