International Journal of Modern Developments in Engineering and Science
Volume 1, Issue 7, July 2022
https://www.ijmdes.com | ISSN (Online): 2583-3138

Personality Prediction Using Social Media Post

Niya Francis[1*], Soumya Sara Koshy[2]
[1]M.Tech. Student, Department of Computer Science and Engineering, Saintgits College of Engineering, Kottayam, India
[2]Assistant Professor, Department of Computer Science and Engineering, Saintgits College of Engineering, Kottayam, India

***Abstract*: Personality is what that differentiate an individual with another. By knowing and understanding an individual's personality, many advantages can be obtained. Together with rapid growth of technology, knowing an individual's personality can be done automatically. Psychology researches suggest that certain personality traits have correlation with linguistic behavior. Supported by the fame of social media, predicting human's personality from their post become possible. Most existing researches have done similar approach in predicting personality from social media. However, focuses on closed vocabulary investigation with English as their language and mostly based on Big Five personality type. As a result, we get Naïve Bayes classifier outperforms the other statistical model with the highest accuracy (80% for I/E and 60% for S/N, T/F, and J/P personality traits) and shows the best performance in terms of speed in classifying the users.**

***Keywords*: Personality prediction, Machine Learning, Naïve bayes.**

1. Introduction

Every person has different hobbies, interests, personality, and other things that define them as a unique person. From all of these differences, we can categorize or classify them to different groups so that it can be useful to improve the effectiveness of advertising, promotion, other marketing purposes, measuring job performance and other functions. Until now, almost all people are using social media. In addition, it is very often that they post something on social media that relates with their interest or hobby which later can be analyzed to know their personality type. This moreover, if the psychological behavior can be accurate and precise for every person that is active in social media, we can do some kind of "foreseeing" what the person will like even though they have not posted it in the internet, which is very advantageous foe marketing and other purposes. By combining all of the information from the social media like Twitter, we can classify people's personality without requiring them to fill a long and time-consuming survey.

The main objective of this research is to develop a: machine learning model and provide a word analysis for each personality traits that might be useful for future researches related to automated personality classification. Naïve Bayes classifier and SVM will be used for the machine learning approach due to its simplicity yet powerful accuracy compared to other algorithms. The grammatic rule is chosen because of its ability to detect context and remove some unnecessary words in a sentence.

By having an automated personality classifier, it can be used to do an effective and targeted marketing for each person that has certain personality, as different personality has different taste or interest in some advertisement or promotion. Or even more, a psychological test can also be done in an employment for candidates so that, HR department can do faster psychological test for the candidates and also job performance measure for the employees.

Table 1
Literature Survey of the research topic

Research	Objective	Approach	Performance
Bharadwaj et al. (2018)	Personality prediction from online text	SVM, Neural Net and Naïve Bayes	SVM with all feature vectors achieved best accuracy across all dimensions of MBTI
Gjurkovi ć and Šnajder (2018)	Personality classification of Reddit user's posts.	SVM, Logistic Regression and MLP with linguistic features	MLP using all linguistic features together outperform across all MBTI dimensions
Plank and Hovy (2015)	Personality and gender prediction from tweets.	Logistic regression Model and Binary word n-gram is used as a feature selection	Accuracy for personality prediction: I/E = 72.5% S/N = 77.5% T/F = 61.2 % J/P = 55.4%
Pratama and Sarno (2015)	To recognize personality using Big-5 personality model from tweets posted in English and Indonesian language	Supervised • KNN • NB • SVM	Accuracy KNN = 58% NB = 60% SVM = 59%
Kaur and Gosain (2018)	Comparing of oversampling and under sampling techniques for imbalance dataset.	Decision tree algorithm C4.5 is used.	Result of Oversampling method (SMOTE) is better than under sampling technique
Chaudhary et al. (2018)	To predict user's personality from the online text using MBTI model	Supervised learning methodology namely Naïve Bayes, SVM, LR and Random	Accuracy NB = 55.89% LR = 66.59% SVM=65.44

*Corresponding author: niya.francis2020@gmail.com

2. Methodology

i. Data acquisition and re-sampling
ii. Pre-Processing and feature selection
iii. Text-based Personality classification using MBTI model
iv. Applying Supervised Machine Learning Algorithms
v. Comparing the efficiency of each algorithm with other classifiers
vi. Applying different evaluation metrics

The data is divided into Training, Testing and Validation dataset. Mostly two datasets are required, one for building the model while the other dataset is needed to measure the performance of the model. Here training and validation are used for building the model, while Testing step is used to measure the performance of the proposed model. The publicly available benchmark dataset is acquired from Kaggle. This data set is comprised of 8675 rows, where every row represents a unique user. Each user's last 50 social media posts are included along with that user's MBTI personality type (e.g. ENTP, ISJF). As a result, a labelled data set comprising of a total 422845 records, is obtained in the form of excerpt of text along with user's MBTI type. At the point when the dataset is divided into training data, validating data and testing data, it utilizes just a portion of dataset and it is clear that training on minor data instances the model won't behave better and overrate the testing error rate of algorithm to set on the whole dataset. Different pre-processing techniques and Classification techniques are exploited, for more exploration of the personality from text. These techniques include tokenization, removal of URLs, User mentions and Hash tag, word stemming, stop words elimination.

Preprocessing: The following preprocessing steps on mbti_kaggle dataset are applied before classification, acquired from the work.

a) Tokenization: Tokenization is the procedure where words are divided into the small fractions of text. For this reason, Python-based NLTK tokenizer is utilized.

b) Dropping Stop Word: Stop words don't reveal any idea or information. A python code is executed to take out these words utilizing a pre-defined words inventory. For instance, "the", "is", "an" and so on are called stop words.

c) Word stemming: It is a text normalization technique. Word stemming is used to reduce the inflection in words to their root form. Stem words are produced by eliminating the pre-fix or suffix used with root words.

Classification: In this proposed work, supervised learning approach is used for personality prediction. The model will take snippet of post or text as an input and will predict and produce personality trait (I-E, N-S, T-F, J-P) according to the scanned text. Mayers-Briggs Type Indicator is used for classification and prediction [4]. This model categorizes an individual into 16 different personality types based on four dimensions.

(i) Attitude: Extroversion vs Introversion: This dimension defines that how an individual focuses their energy and attention, whether get motivated externally from other people's judgement and perception, or motivated by their inner thoughts.

(ii) Information: Sensing vs. Intuition (S/N): This aspect illustrates that how people perceive information and observant(S), relying on their five senses and solid observation, while intuitive type individuals prefer creativity over constancy and believe in their guts

(iii) Decision: Thinking vs. Feeling (T/F): A person with Thinking aspect, always exhibit logical behavior in their decisions, while feeling individuals are empathic and give priority to emotions over logic

(iv) Tactics: Judging vs. Perceiving (J/P): This dichotomy describes an individual approach towards work, decision-making and planning. Judging ones are highly organized in their thoughts. They prefer planning over spontaneity. Perceiving individuals have spontaneous and instinctive nature. They keep all their options open and good at improvising opportunities.

3. Results

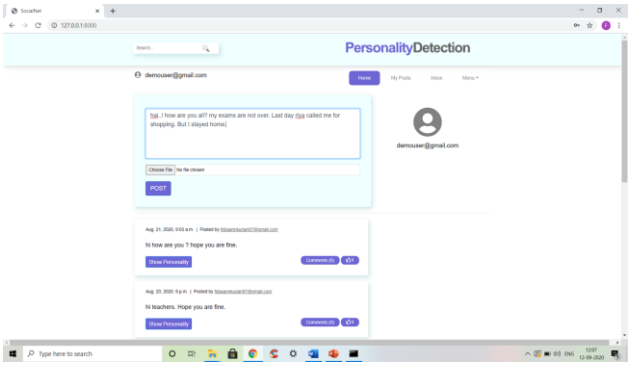

Fig. 1. User home page

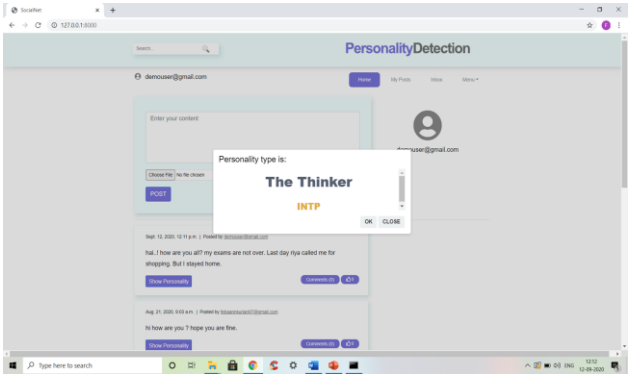

Fig. 2. User view personality

4. Conclusion

The new system has overcome most of the limitations of the existing system and works according to the design specification given. The developed systems dispense the problem and meet the needs of by providing reliable and comprehensive information. All the requirements projected by the user have been met by the system. The newly developed system consumes less processing time and all the details are updated and processed immediately. Since the screen provides online help messages and is very user- friendly, any user will get familiarized with its usage. Modules are designed to be highly flexible so that any failure requirements can be easily added to the modules without facing many problems. In future as time

passes by, we can modify this project to suit the requirements of the time and future modules can be added to it.

References

[1] A. Furnham and J. Crump, "The myers-briggs type indicator (mbti) and promotion at work," Psychology, vol. 6, no. 12, p. 1510, 2015.

[2] M. D. Back, J. M. Stopfer, S. Vazire, S. Gaddis, S. C. Schmukle, B. Egloff, and S. D. Gosling, "Facebook profiles reflect actual personality, not self-idealization," Psychological science, 2010.

[3] V. Bijalwan, P. Kumari, J. Pascual, and V. B. Semwal, "Machine learning approach for text and document mining,", 2014.

[4] J.-Y. Yoo and D. Yang, "Classification scheme of unstructured text document using tf-idf and naive bayes classifier," 2015.

[5] I. Myers and P. Myers, Gifts differing: Understanding personality type. Nicholas Brealey Publishing, 2010.