

# Comparative Analysis for Personality Prediction by Digital Footprints in Social Media

Mr. R. Valanarasu,

Senior Consultant,

Enterprise Application Services,

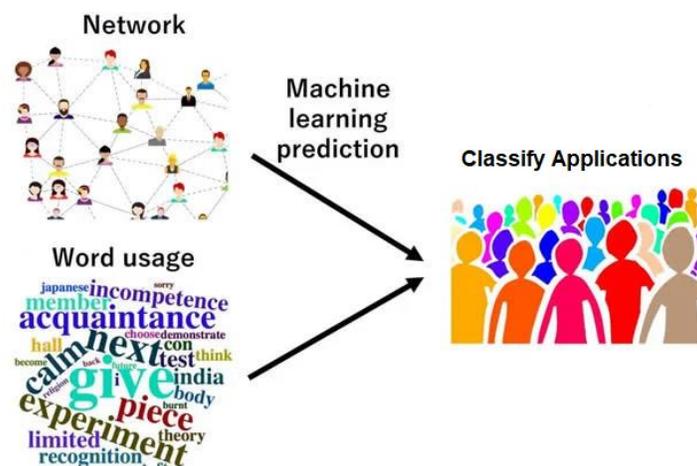
Infosys Ltd.

**Abstract:** The use of social media and leaving a digital footprint has recently increased all around the world. It is being used as a platform for people to communicate their sentiments, emotions, and expectations with their data. The data available in social media are publicly viewable and accessible. Any social media network user's personality is predicted based on their posts and status in order to deliver a better accuracy. In this perspective, the proposed research article proposes novel machine learning methods for predicting the personality of humans based on their social media digital footprints. The proposed model may be reviewed for any job applicant during the times of COVID'19 through online enrolment for any organisation. Previously, the personality prediction methods are failed due to the differing perspectives of recruiters on job applicants. Also, this estimation is modernized and the prediction time is also reduced due to the implementation of the proposed hybrid approach on machine learning prediction. The artificial intelligence based calculation is used for predicting the personality of job applicants or any person. The proposed algorithm is organized with dynamic multi-context information and it also contains the account information of multiple platforms such as Facebook, Twitter, and YouTube. The collection of the various dataset from different social media sites constitute to the increase in the prediction rate of any machine learning algorithm. Therefore, the accuracy of personality prediction is higher than any other existing methods. Despite the fact that a person's logic varies from season to season, the proposed algorithm consistently outperforms other existing and traditional approaches in predicting a person's mentality.

**Keywords:** *Deep learning, Personality prediction*

## 1. INTRODUCTION

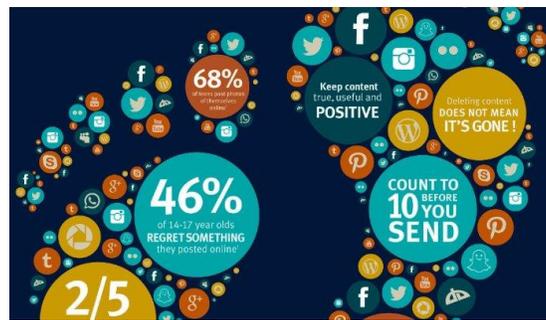
Personality refers to the combination of behaviour and emotions, which have a unique set of qualities depending on the environment and biological causes. This will be varying from person to person depending on many conditions based on their thinking and feelings [1]. The personality quality will construct in nature based on their activities, thinking, feelings, and overall behavior in different conditions. Generally, this activity of any human is based on his nature and knowledge. This will vary from person to person which comes under the study of personality psychology. To measure different behavior for different individuals is that the consistency in many situations with stability [2]. Recently, social sites are increasing abruptly for different communication purposes. People are using those platforms to share their thoughts, expectations, and feelings. Based on their likes and unlike activities, they can be categorized easily. This information is collected as a dataset by social media [3]. Initially, this information is used for banking sectors to identify the person's location and position. The extracted data has been explored year by year in many sectors. Besides, improving the quality of services and products is done by this phenomenon [4]. Figure 1 shows the classification of job applications through word usages in social media.



**Figure 1** Classify the Job Applications based on Personality

Moreover, sentimental analysis is used to detect the emotions are positive or negative of humans on the same topic. From this, researchers are proposed to determine the mental health

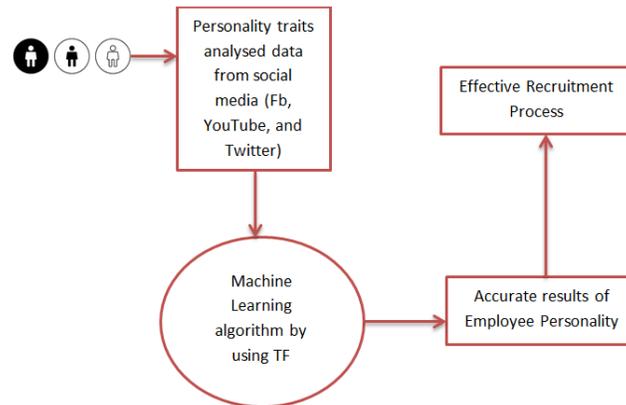
issues, stress level, and behavior of the person based on their activities in social media [5]. Nowadays, the accessing internet and social media are becoming habitual in everyday life. People are using for sharing their thoughts, feelings, expectations, and opinions. The social networking sites are Facebook, Twitter, Instagram, and YouTube which are used to interact with the user asking for content [6]. This can create to start the exchanging data and interact freeform of data production. Also, social media addicted people will publish their thoughts immediately in social networking sites through their status, post sharing, likes and dislikes, photos, videos, and self-description interest. Here, the data is available for social web connections in the huge volume of shared posts [7]. The proposed motivation is to explore the chance of predicting people's mental stress, health, and activities through social media network datasets. Figure 2 shows some digital footprints in social media.



**Figure 2** Digital Footprints in Social Media

Recently, many MNC sectors are recruiting people based on their mental stress, activities in social networking through human resource management. Human resource management is one of the significant sources in the company sector to recruit and making guidelines for the employees. The traditional approach is having a collective series of questions to evaluate their behavior of the candidate but this method is time-consuming and a failure for some sort of people. If they can act nicely no one knows about him in front of the interview panel. However, it has been updated recently based on people's thoughts, knowledge, and trend. This digital footprint gives a more accurate answer than the traditional method. The author also believes in the recent follow-up of digital footprints on the person might be a good judgment and provide

accurate answers within a short period of time [8]. This judgment is more accurate than those user's friends and even wife or husband prediction.



**Figure 3** Simplified work flow of proposed method

Figure 3 shows a simple workflow graph of proposed method. There is a necessity to find more accurate answers through many algorithms and finding a variety of approaches based on the model predictions can be explored in the research work. There are different types of prediction mechanisms and limited feature extraction techniques based on the availability of the dataset. The proposed research work includes the algorithm with results obtained as well.

## 2. ORGANIZATION OF THE RESEARCH

The proposed research article is arranged as follows; section 3 provides preliminary work on personality prediction by using various algorithms; section 4 discusses various algorithms to predict the personality from social media users. The obtained results and discussions are described in section 5. Section 6 concludes the overall research work and also includes a discussion on its future extension.

## 3. PRELIMINARIES

Maite et al proposed a technique to predict personality by incorporating the machine learning algorithm. They predicted any person's personality from the Twitter database. They concentrated on many languages and this paper is mainly focused on English language only.

They conducted five models for testing through word embedding, which is considered as short input data. They have used glove representation in vector method and it is used to compute the word embedding procedure. They have implemented CNN methods, which were used to predict the personality by using various filters of convolution layers. The pooling layers were applied to merge all the output obtained from various layers present in the structure. Finally, the activation function was used to efficiently select the output [9]. Salem et al focus on the personality prediction of Arabic and Egypt based Twitter users. They collected a database from the Arabic Twitter users. They created a questionnaire with 5 choices like a MCQ type, which can be used to predict the users' psychological order. These questions were framed in Arabic language and they can also be translated into English. They have collected both personalities by determining questionnaire answer and their post as well. They have done preprocessing procedure for the news feed post to remove the noisy pattern from it. Also, they have incorporated Arabic translator in their algorithm for performing any normalization procedure. They suggested many machine learning algorithms named decision trees; support vector machine (SVM), and naïve Bayes methods [10]. M Hassanein et al focuses on predicting the human personality by using semantics. Here, they collected the dataset from different social media. Also, various machine learning algorithms such as support vector regression, and feed-forward neural networks are implemented to predict the personality. The word in the post can be counted and represented in the vector space model. This can be measured for analyzing similarity measures in order to predict the personality by WordNet Dataset [11]. Thread et al proposed the personality prediction process by using the social media platform. Here, three different machine learning algorithms are implemented to predict the personality from the given data. The gradient boosting method, XGB model, and Pearson correlation methods are used to predict the personality. They predict the personality plan, which is based on the marketing strategies. In the other hand, it can be extended to improve relations with the users. They have used personality datasets and branded page datasets, which are used for performing feature extraction with the machine learning procedure [12].

X sun et al introduce a new approach by integrating bidirectional LSTM network with CNN to predict the personality of users. They are mainly focused on big five models to compare

their proposed work with a long text dataset of essays. They have also focused on the structure of text-based important features. Besides, they have used a glove algorithm, which is used for embedding the word. They contained self-loop and Recursive neural networks as well as to extract the features from essays. They concentrated on latent sentence group procedure to classify several sentences between the people. This study is used for various latent features by the CNN model. Finally, the softmax activation function is used to accurately classify a maximum number of possibilities. They have proved their algorithm as superior method than other big models [13]. Vaidhya et al analyze the personality functions for many social media users by using their status post. They are examined with big five personality models for prediction. Also, very famous datasets named MyPersonality dataset are used by many users. The post can be pre-processed due to noisy information such as links and others. They appended a spelling correction tool, which is used to correct spelling mistake in the post with their algorithm to obtain better accuracy than other algorithms. They are analyzed with the principal component analysis method [14].

The prediction process is collected through the label distribution learning method. They collected various datasets from microblogging sites; they considered status etc. They have calculated the overall personality scores with 44 questionnaires. They extracted the features from the post by considering the basic information like gender, name, etc. The content feature is always a very dynamic feature. It has included the linguistic, psychological features, etc. They have conducted many examinations such as linear regression, label distribution with a support vector machine classifier and machine-learning algorithm to predict the personality from their post on social media [15].

T. Yo et al proposed the prediction process based on age, gender, and occupation details obtained from the Twitter dataset. Metab tool was used to collect the words. Also, word 2vec model is used to compute the word embedding. They compare various machine learning methods such as random forest, KNN, AdaBoost, Full connection neural network. The prediction of gender and profession will provide better accurate results on the final stage of predicting the human personality [16].

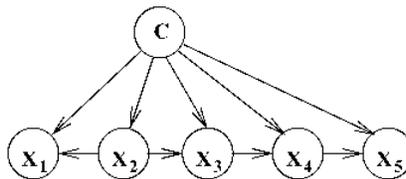
## 4. METHODOLOGIES

A comparative study has been performed on the computation analyses of various prediction algorithms used for identifying human mental stress and activities. This personality prediction is treated as a multi-label prediction task to achieve more accuracy. A hybrid analysis technique is proposed for performing personality prediction by using their digital footprints in social media.

### 4.1 Machine Learning Techniques

#### 4.1.1 Naïve Bayes Model

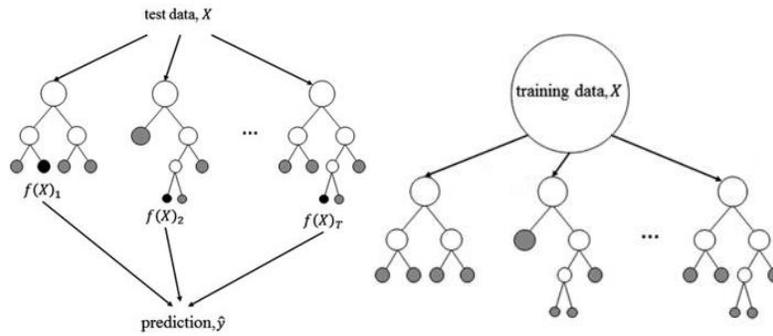
The Naïve Bayes algorithm is a probabilistic method that is used to predict the events with a huge amount of datasets. The likelihood theory is proved with Bayes's hypothesis for resulting in limited learning probability [17]. The primary motivation of naïve bayes classifier is its simple database, which is used for establishing standardization in the outputs. Figure 4 shows the simplified model of the Naïve Bayes algorithm.



**Figure 4** Simplified Structure of Naïve Bayes Model

#### 4.1.2 Random Forest (RF)

This method is a controlled learning algorithm, which has a primary moment of scope. This algorithm is working well in grouping the events and revert issues. This can be predicted by using the discrete makings, which are used to recognize the acquired results through properties. The classification is faster to manage the event, which is not present in the dataset [18] [19]. Figure 5 shows the training and testing data of prediction flow.

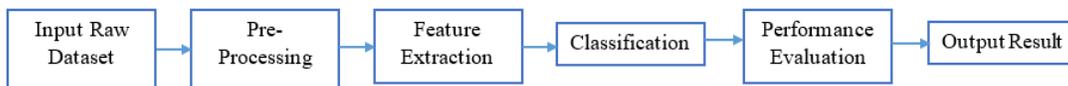


**Figure 5** Training and Testing Structure through Tree Model of RF

### 4.2 Hybrid Proposed System

The combination of the Naïve Bayes algorithm and SVM classifier construction is implemented in the proposed model to predict the personality of the person with high accuracy by using their digital footprints in social media. The average value of the class and corrected variance can be updated for prediction based on the Bessel function [20] [21]. The block diagram of the proposed system is shown in figure 6. The normal distribution is taking place in class after updating,

$$P(x = v/c_n) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{\left[\frac{(v-\mu_n^2)}{2\sigma_n^2}\right]}$$



**Figure 6** Block diagram of Overview of Proposed Technique

The raw input datasets are preprocessed for performing feature extraction in various social media platforms like Facebook (Fb), Twitter and YouTube, etc. The extracted feature details are structured through the Naïve Bayes algorithm. The customized user contact is comprised of varying media content, status update history, numerous inactive factors, and users’ active posts with sharing thoughts. The Naïve Bayes algorithm is used to predict the events in a

huge amount of datasets [22]. The final stage of the proposed framework is completed with an SVM classifier for obtaining higher accuracy results.

The pre-trained datasets will provide a good classification rate by using SVM classifier for any predicted events. The data points are separated based on their events/ similarities. Here, the logistic hyper-plane is used to acquire a higher accuracy rate [23] [24]. The kernel vector is estimated to approach the real datasets. The magnification of the attributes involved in the given inputs by SVM is defined as,

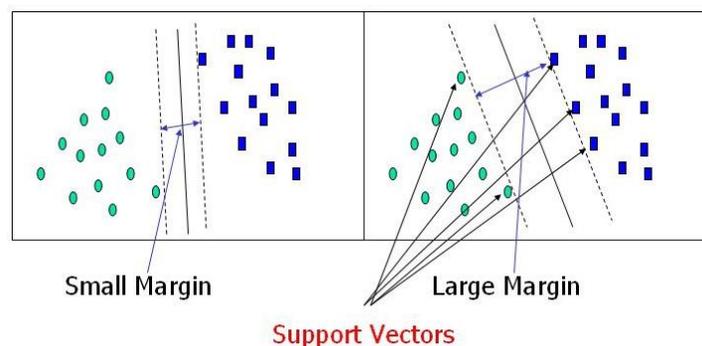
$$= \frac{1}{n} \sum_{k=1}^n \varphi_k + \lambda \|w\|^2$$

These kernel factors are transforming from linear to nonlinear classification, which can provide better accuracy. This nonlinearity is mapping from small margin to large margin scale. Finally, the error minimization in the classification time is defined as,

$$e(f) = \frac{1}{n} \sum_{i=1}^n l(\hat{u}, f(e_i))$$

Where,  $\varepsilon_i$  is an empirical variable.

Any organizations with a recruiting process for human resources are essentially observed on the employee. Figure 7 shows a small and large margin support vectors graph.

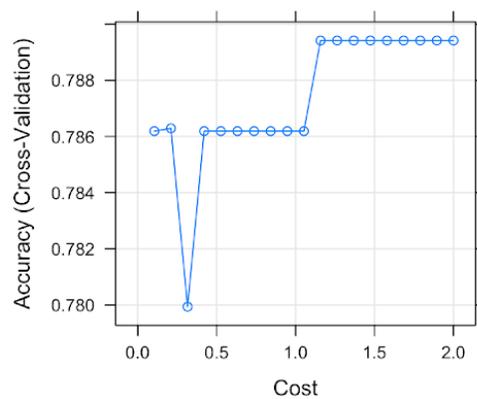


**Figure 7** Small margin and large margin support vector

## 5. RESULTS & DISCUSSION

### *Various Data Stream for Personality Prediction*

Recently, the information on Facebook is utilized to analyze the character of a person with the available personality analysis datasets. The character element can be interpreted based on the content of the information. The Twitter datasets are used frequently in personality prediction ventures. Due to the large set of users in Twitter, the prediction accuracy is increased based on strong character acknowledgment. Another examination of video datasets is created by YouTube social media that remarks user's enthusiasm characters are being noted. This research work has considered the various ways of including literary, media content, inert factors, and frequent updating status, which is most important to classify the effects. Figure 8 shows the graph between accuracy vs. cost.



**Figure 8** Accuracy vs. Cost

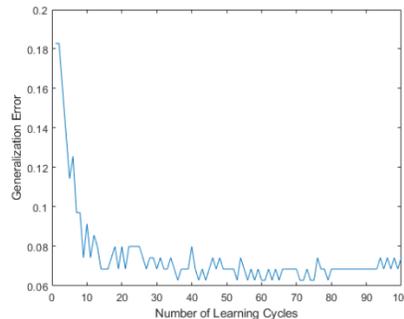
This test can forecast the social media users' information and their mentality. The potential can be increased in numerous computational applications. The data has been sorted in the tensor flow library to reduce the searching time in algorithm format. It is analyzed and predicted by tensor flow python coding. This combined hybrid algorithm has provided good intelligent and active employees to an organization. The following formulas are used for investigation.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

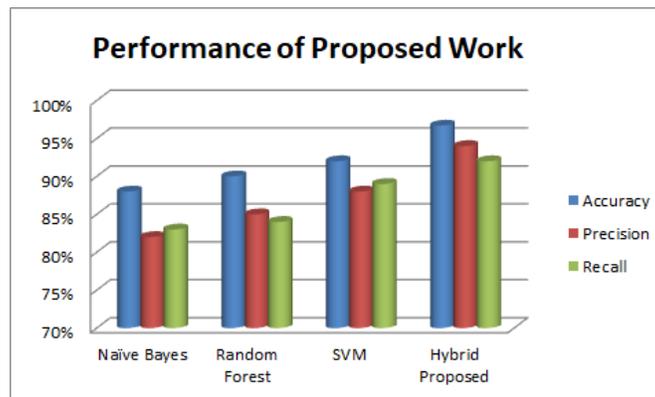
$$Recall = \frac{TP}{TP + FN}$$

$$F - Measure = \frac{2.TP}{2.TP + FP + FN}$$



**Figure 9** Error Minimization in Classifier

This accuracy measure has been performed for a huge amount of dataset and it's around 80%. Figure 9 shows the error minimization in our proposed classifier after naïve Bayes probability prediction.



**Figure 10** Overall Performance of Proposed Work

This computational character investigation is examining the cutting edge via the changed arrangement of social media addictions. The social media database is used for character

expectations and tasks. Figure 10 shows the overall performance by illustrating a graph chart on the proposed work. Here, False negative rate is missing some individual examination results.

**Table 1** Overall Performance of Machine Learning Methods

S.NO	METHODS	ACCURACY	PRECISION	RECALL	False Negative Rate
1	Naïve Bayes	88%	82%	83%	9.4%
2	Random Forest	90%	85%	84%	7.1%
3	SVM	92%	88%	89%	6.53%
4	Hybrid Proposed (NB + SVM)	96.7%	94%	92%	2.5%

It is almost negligible with the hybrid proposed framework. But the single type classifiers are facing a problem of more classification error and computation time also. Table 1 shows the overall performance of the proposed framework.

## 6. CONCLUSION

Thus, the proposed machine learning model has been developed to predict the personality with comparatively higher accuracy. Online social media has been comprised of many emotional and personally descriptive contents to reveal. Besides, the language translators are also used in the proposed algorithm for analyzing all non-English content extraction. The social media networks are allowing the users to use local language to reach final boundary people. This combination of Naïve Bayes and SVM methods has provided good prediction and classification performance while analyzing the personality. The naïve Bayes is probabilistic and it provides a better prediction rate than all other machine learning algorithms. Besides, SVM provides higher classification accuracy rate and minimum classification error rate. The hybrid version of these machine learning algorithms provide a good prediction rate in the personality prediction paradigm by leveraging higher accuracy and minimum relative error in the classification. From the text, the characters and feelings are interpreted with the help of hybrid classification. The authors strongly believe that, the proposed hybrid algorithm can predict the personality of the

person with comparatively higher accuracy. This will be a benefit for human resources present in various sectors of ICT industries on their recruitment process. They can predict the personality of any job applicant with higher accuracy by using the proposed algorithm. In this research article, the relative investigation on computational character acknowledgment has settled and compared with ground truth information obtained from the users. The ground truth datasets are created for Facebook, Twitter, and YouTube by using the manual reasoning method. In the future, the following issues will be addressed; if the job applicants are non-social media networking users, then the proposed framework cannot be used. There will not be any other option to predict the personality. These are the drawbacks observed in the proposed system. In the future, the proposed algorithm can be incorporated with traditional questionnaire-based personality predictors, which can provide high accuracy and better prediction.

## REFERENCES

- [1] M. K. Hayat, A. Daud, A. A. Alshdadi, A. Banjar, R. A. Abbasi, Y. Bao, and H. Dawood, "Towards Deep Learning Prospects: Insights for Social Media Analytics," *IEEE Access*, vol. 7, pp. 36958–36979, 2019.
- [2] Victor Zhou, *Machine Learning for Beginners: An Introduction to Neural Networks*, Towards Data Science, 05-Mar-2019.
- [3] H. Thilakarathne, *Artificial Neural Networks with Net# in Azure ML Studio*, NaadiSpeaks, 08-Nov-2017.
- [4] R. Wald, T. Khoshgoftaar, and C. Sumner, "Machine prediction of personality from Facebook profiles," *2012 IEEE 13th International Conference on Information Reuse and Integration (IRI)*, 2012.
- [5] N. Majumder, S. Poria, A. Gelbukh, and E. Cambria, "Deep Learning- Based Document Modeling for Personality Detection from Text," *IEEE Intelligent Systems*, vol. 32, no. 2, pp. 74–79, 2017.
- [6] H. Zheng and C. Wu, "Predicting Personality Using Facebook Status Based on Semi-supervised Learning," *Proceedings of the 2019 11th International Conference on Machine Learning and Computing - ICMLC 19*, 2019.

- [7] Laleh and R. Shahram, Analyzing Facebook Activities for Personality Recognition, 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA), 2017
- [8] B. Y. Pratama and R. Sarno, Personality classification based on Twitter text using Naive Bayes, KNN and SVM, 2015 International Conference on Data and Software Engineering (ICoDSE), 2015.
- [9] Maite Giménez, R. Paredes and R. Paolo, "Personality Recognition Using Convolutional Neural Networks", Springer Nature Switzerland AG pp- 313-323, 2018
- [10] M. S. Salem, S. S. Ismail, and M. Aref, "Personality Traits for Egyptian Twitter Users dataset," Proceedings of the 2019 8th International Conference on Software and Information Engineering - ICSIE 19, 2019.
- [11] M. Hassanein, W. Hussein, S. Rady, and T. F. Gharib, "Predicting Personality Traits from Social Media using Text Semantics," 2018 13th International Conference on Computer Engineering and Systems (ICCES), 2018.
- [12] R. B. Tareaf, P. Berger, P. Hennig, and C. Meinel, "Personality Exploration System for Online Social Networks: Facebook Brands As a Use Case," 2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI), 2018
- [13] X. Sun, B. Liu, J. Cao, J. Luo, and X. Shen, "Who Am I? Personality Detection Based on Deep Learning for Texts," 2018 IEEE International Conference on Communications (ICC), 2018.
- [14] M. Vaidhya, B. Shrestha, B. Sainju, K. Khaniya, and A. Shakya, "Personality Traits Analysis from Facebook Data", 21st International Computer Science and Engineering Conference (ICSEC), 2017.
- [15] D. Xue, Z. Hong, S. Guo, L. Gao, L. Wu, J. Zheng, and N. Zhao, "Personality Recognition on Social Media With Label Distribution Learning," IEEE Access, vol. 5, pp. 13478–13488, 2017.
- [16] T. Yo and K. Sasahara, "Inference of personal attributes from tweets using machine learning," 2017 IEEE International Conference on Big Data (Big Data), 2017.

- [17] Xiong, Qian, et al. "Privacy-Friendly Personality Recognition in Social Media: A Case Study of Chinese WeChat Users." In International Conference on Applications and Techniques in Cyber Security and Intelligence, 2019.
- [18] Kim, Carolyn, and Karen Freberg. "The state of social media curriculum: exploring professional expectations of pedagogy and practices to equip the next generation of professionals." (2017).
- [19] Howlader, et al. "Predicting facebook-users' personality based on status and linguistic features via flexible regression analysis techniques." 2018.
- [20] Gonzalez, Roberto J. "Hacking the citizenry Personality profiling, 'big data' and the election of Donald Trump." 2017.
- [21] Iannelli, Laura, et al. "Facebook digital traces for survey research: Assessing the efficiency and effectiveness of a Facebook Ad-based procedure for recruiting online survey respondents in niche and difficult-to-reach populations. (2018)
- [22] Buettner, Ricardo. "Personality as a predictor of Business Social Media Usage: an Empirical Investigation of Xing Usage Patterns." 2016.
- [23] Azucar, et al.. "Predicting the Big 5 personality traits from digital footprints on social media: A meta-analysis." Personality and individual differences 2018.
- [24] Obschonka, et al. "Using digital footprints in entrepreneurship research: A Twitter-based personality analysis of superstar entrepreneurs and managers." (2017).