Software Proposal Document for
Personality Traits Detection Based on Facebook Using Machine learning

Amal Khaled, Karim Ali, Mohamed Sabry, Mohamed Akram
Supervised by: Eng. Lobna Shahen, Prof. Dr. Essam Halim Houssein

December 14, 2021

Proposal Version	**Date**	**Reason for Change**
1.0	9-November-2021	Proposal First version's specifications are defined
1.1	10-November-2021	Scope updated

Table 1: Document version history

GitHub: https://github.com/Amal-Safwat/Automatic-Recognition-of-Personality-Traits-on-Facebook.git

Abstract

The use of social media is rapidly growing. Various information's are shared widely through social media, such as Facebook. Information about users, as well as what they conveyed through status updates, is extremely valuable for behavioral learning and human psychology research. Similar studies have been undertaken in this subject, and it grows continually till now. This research aims to develop a system that can predict a person's personality based on information from Facebook accounts. We propose an approach that intends to facilitate this line of research by directly predicting the Myers Briggs Type Indicator (MBTI) personality analysis from user's Social Network Sites (SNSs) behaviors. Comparing to the conventional inventory-based psychological analysis. We demonstrate via experimental studies that users' personalities can be predicted with reasonable precision based on their online behaviors.

1 Introduction

1.1 Background

The analyzation of human personality can be mad through the merging of some different traits that each of us has either of them. The Myers-Briggs Type Indicator (MBTI) is an international personality test, that aims to differentiate psychologically differences between how the people perceive the world and make their decisions in every situation [1]. This test was made for normal people to differentiate between people based on these traits as shown in Figure 1.

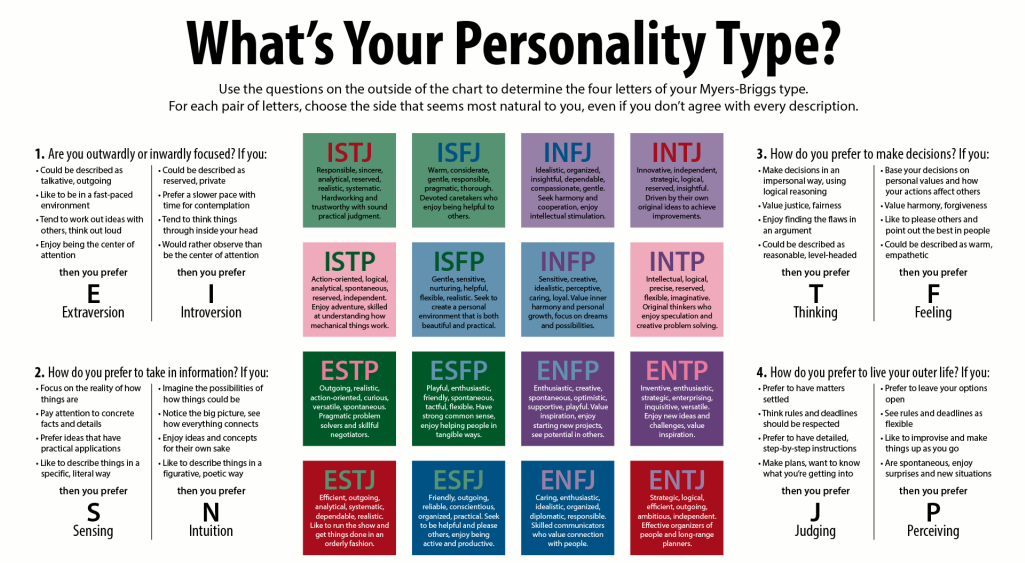


Figure 1: MBTI Personality test [7].

The four personality classes on the MBTI scale provide a wide range of personality classification options. They are divided further into sixteen different personality types as a binary combination of 4 digits. The four MBTI classes are the following [2] [7]:

Extroverted vs Introverted: In this class we are going to see if the person is outgoing, talkative and loves to meet new people or the contrary.

Sense vs Intuition: In this class we see indication of how the person perceives information. Some people use logic about what they are sure of while other ones use their intuition and instincts and follow them.

Thinker vs Feeler: In this class we differentiate between people's decision making processes as we humans are divided into two categories. Some of us makes decisions based on logic and other ones makes it based on their feelings and emotions.

Judge vs Perceive: In this class we differentiate between people from their orientation to the outer world. Some are always organized and settled while others are spontaneous most of the time.

In this project we aim to help people understand themselves more through knowing their personality types. Also we aim to help psychologists to treat patients more efficiently through helping them to know their personalities before they visit them.

1.2 Motivation

1.2.1 Academic

The problem of trying to understand the human brain is a very old problem since the days of the great philosophers of Greece [3] and everyone is trying to add more to it in order to be more accurate while doing so, too many solutions has been brought to the table for understanding people and their behaviors of which the most successful are the personality tests like 16 personalities test and Facebook AI that studies people's thinking and behaviors through what they do everyday, but every day someone try to come up with an idea that can improve the accuracy or enhance human understanding in a different way, but no one ever reaches a limit for that problem.

1.2.2 Business

This project would be needed by anyone who works in the psychological field as it would improve their work if they know the personality of their patients before they come to them, also with some improvements it would be used by marketing companies too. It can also be used by people who needs job recommendation based on their personality and their strengths and weaknesses.

1.3 Problem Statement

1. The human errors are huge in the process of hiring people, we always have no time, patience and of course not enough knowledge to be able to judge which candidates are the best for the job. Knowing someone's personality, preferences and what career they would be best at would minimize these human errors.

2. In the interviews people spend most of the time trying to identify the candidate personality, which is a problem that costs a lot of time. Through knowing the candidate's personality through their social media this time would be minimized which would decrease the cost to the minimum and make the interview job easier for everyone.

3. The candidate may not be sure about which career will be the best fit for his/her personality and which environment will be more suitable for his/her traits. So he/she may need job recommendation that matches his/her strengths and weaknesses.

2 Project Description

According to bring your Personality Dimension "MBTI" is one of the best personality measures to serve as a research model and is considered good in measuring one's personality [14]. When it comes to dealing

with data imbalance Natural Language Processing(NLP)is among the best topic in the field of data science. Companies are putting tons of money into research in this field. The help of natural language processing, recruiters can find the right candidate with much ease.but before we do anything we need to ask the user permission to access his/her profile on social media to get all the data and posts needed from his/her profile. Lately the NLP method is considered as an advanced approach that will help us in solving this issue as shown in Figure 2. So, the personality assessment using social media process by analyzing the text extracted from the candidates' posts is very useful that can be used to suggest general job roles suitable for the candidates, and know the suitable job environment for his personality [3]. And this helps in saving time for the hiring process as the company will have an overall idea of the candidate's personality and the job role he/she is suitable for.As there is a strong interconnection between a user's personality and their online behavior on social networks has been observed in various studies and experiments. Since social media posts made by users are reflective of their personality traits, they can be analyzed to predict the personality of a person as shown in Figure 3. First is the approach at the data level, the second is the approach based on the algorithm level, and the last is by using a combination of learning methods by combining several classification methods [14].Not only but also , adding to our project to extract objects from the images that users share on their social media to increase the accuracy of choosing the personality trait for the user.

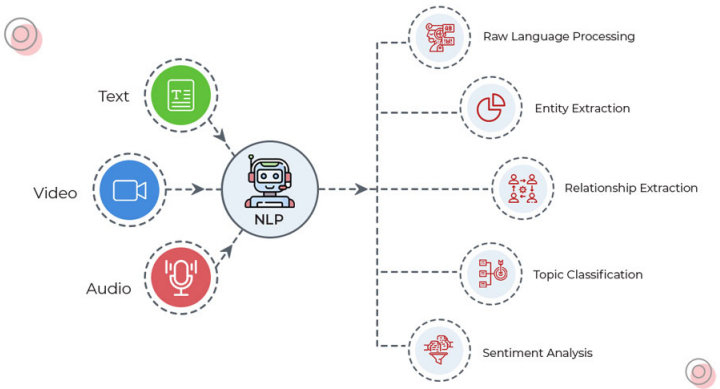


Figure 2: Natural language processing.

ISTJ factual practical organized steadfast	**ISFJ** detailed traditional service-minded devoted	**INFJ** committed creative determined idealistic	**INTJ** independent visionary original global
ISTP logical realistic adventurous self-determined	**ISFP** caring adaptable gentle harmonious	**INFP** compassionate original creative empathetic	**INTP** independent theoretical analytical reserved
ESTP activity-oriented versatile pragmatic outgoing	**ESFP** enthusiastic friendly cooperative tolerant	**ENFP** creative versatile perceptive imaginative	**ENTP** enterprising outspoken challenging resourceful
ESTJ logical systematic organized conscientious	**ESFJ** thorough responsible detailed traditional	**ENFJ** loyal verbal energetic congenial	**ENTJ** logical strategic fair straightforward

Figure 3: 16 Personality Traits test.

2.1 Objectives

1. The project has two main objectives the first one, aims to detect person's personality through text written by him, we intend to use machine learning and deep learning models to build a classifier that will take posts as input and produce as output a prediction of the MBTI personality type.

2. The second object aims to create a model where the user will be provided with career guidance which matches his personality. For example: if a user has the ability to speak well and is able to convince the opposite person. So, this user will be good in the marketing field.

3. Preform the accuracy.

2.2 Scope

1. Understand human behaviour through knowing their personality types.

2. To help psychologists in understanding their patients before even they come to them.
3. To give people better understanding for which career paths they should take specially for high school students.
4. Detect Object from Image

2.3 Project Overview

The concept of the system in this study is shown in Figure 4.

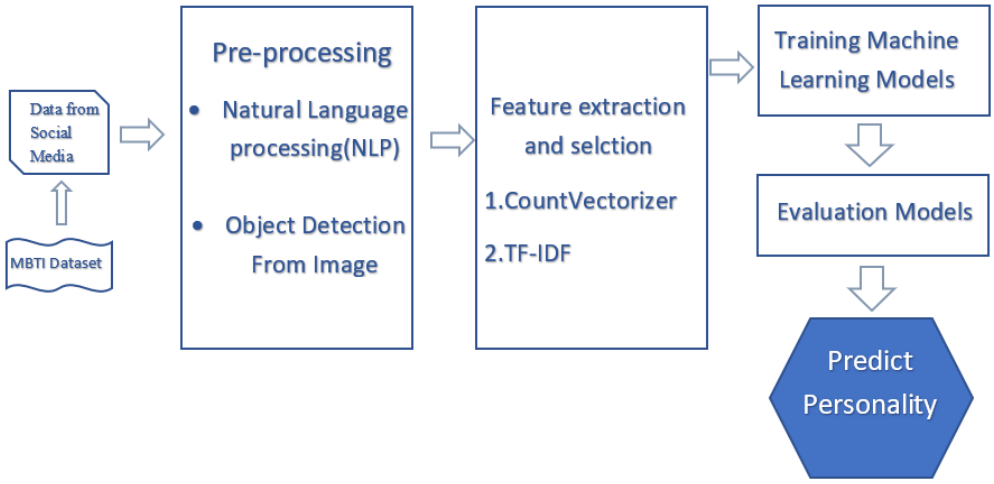


Figure 4: Project Overview.

The system steps of this study are as follows:

1-Data Extraction: This phase involves identifying and collecting data sets. Facebook has been the major source for social media data.

2-Pre-processing Steps:

- **Natural Language Pre-processing:** Text Pre-processing is traditionally an important step for natural language processing (NLP) tasks. It transforms text into a more digestible form so that machine learning algorithms can perform better.

1. Normalize: Changing all the capital letters into lowercase.
2. Tokenization: A common task in Natural Language Processing (NLP). It is a way of separating a piece of text into smaller units called tokens.
3. Filerting: Remove words that are meaningless which are stop words.

4. Stemming & Lemmatization: The objective of stemming is to reduce related words to the same stem even if the stem is not a dictionary word. for this phase we used WordNet lemmatizer as in this phase we translate language to English.
5. Removing URLs / Hashtags / Mentions.

- **Object Detection From Image:**

1. Define Labels to be detected in images.
2. label the images either manually on a moderately sized dataset or using an object detection model that will detect the different objects automatically.
3. create label map that will give an ID to each label.
4. Create records for training and testing then train the model.
5. evaluate the model's precision and recall it.
6. Load model from checkpoint then structure it to be used in implementation independently.

3-Feature Extraction and selection: Methods of feature extraction are:

- countVectorizer: We used to transform a given text into a vector on the basis of the frequency (count) of each word that occurs in the entire text.
- TF-IDF: Means Term Frequency - Inverse Document Frequency. This is a statistic that is based on the frequency of a word in the corpus but it also provides a numerical representation of how important a word is for statistical analysis.

4-Training Machine Learning Models In this phase we will apply some machine learning algorithms to choose the one that gives best accuracy:

- Naïve Bayes
- Support vector machine(SVM)
- Recurrent neural network (RNN)
- k-nearest neighbor (kNN)
- XGBoost

5- Evaluate models: Evaluating results from personality model.

2.4 Stakeholder

2.4.1 Internal

The development team are:

- Amal Khaled is the team leader
- Karim Ali
- Mohamed Sabry
- Mohamed Akram

2.4.2 External

Our system targets the companies, it can be used by the HR during the hiring process to be able to choose the right candidate. The employers in different fields and people who can't find suitable jobs for themselves. It can also be used by people who needs job recommendation based on their personality and their strengths and weaknesses.

3 Similar System

3.1 Academic

Although researches on Personality trait recognition are somewhat scarce, there are many related works in this domain of research, the following researches are the ones with the most similarity to our project:

Salem et al, M 2019[10] This research was done in Egypt where the main objective of the project was to reach an artificial intelligent system that can make a use of the data extracted from the social media users to predict their personality traits in Arabic. The researchers set the baseline which will be held for future comparisons for improvements as it was one of the first projects to apply this idea to the Arabic language. The dataset collection was done by two methods: The first was a questionnaire created by psychologists that was published on social media platforms, The second method was done by a psychologist that would observe the subject's social media account which was more accurate. The researchers mainly relied on three learning algorithms which are KNN, DT then SVM algorithms. These algorithms resulted in varying accuracies for each model. However, their average of the best result in each trait is 0.59 and 0.33 in binary and multiclass representation respectively. The main points of criticism are the unbalanced dataset used by the researchers and the very volatile and unreliable accuracies they achieved using multiple models.

Rahman et al, M.S 2019[9] This research is more focused towards the applications of personality traits recognition and approached the problem of natural language processing from a practical view. The researchers developed a large number of features combined with a bigger database to illustrate their effects on the accuracy of the project. The dataset they used was developed by observation of 2,468 people, this large dataset tends to be more accurate for practical applications. The highest average accuracy achieved by using various models is 49.07%, their result concluded that using CNN with the leaky ReLU function produced the best results. A way to improve upon this project is using more models and vary the functions between them to reach the optimal accuracy with the best model.

Majumder et al, E.2017[4] This paper approached this project with a binary classification of the traits to solve the same problem as in the aforementioned research [9]. The authors used multiple models with the same architecture to extract personality traits from essays using a convolutional neural network. This research is concerned with achieving a higher accuracy using the mentioned binary classification. The dataset used is almost the same with 2,467 subjects in the dataset. The project had an average accuracy of 58.09%(EXT), 59.38(NEU), 56.71%(AGR), 57.30%(CON) and 62.68%(OPN) for each of the big-five traits respectively. The authors' approach was solely centered around the binary classification of the traits, which did not increase the accuracy by a considerable amount.

Hassanein et al, T.F.2018[6] The authors used a similar approach to the binary classification method[4], but instead of binary classification they classified the traits to vectors where the dataset input can lie on a spectrum rather than a positive-negative value. This approach resulted in a higher accuracy than the previous papers[4][9] where they used text semantics to measure the similarity between the user's text and the words used for comparison. The dataset used was collected from twitter with 81 different features to train the machine learning model for classification, they also included user profile data to aid in the classification. The highest average accuracy from their models was 64% which is fairly high in comparison with the other

papers. The points of criticism on this research lie on their inconsistent results in most models, where they were giving accuracies between 47% and 57%.

Yang, H.C., Lee, C.H. and Yeh, C.Y. 2020[[[11]] The authors of this paper went for the same idea as the ones mentioned previously, they used text mining and a classification model with the goal of getting a higher precision than previous attempts on the dataset used, which they achieved using a scoring method that gives each trait a score from 1-7 in the classification. They used the my Personality dataset which collected 250 users from Facebook with 9917 status updates with the traits identified in the dataset as well. The researchers achieved an average mean square error of 0.107 which was slightly better than the previous attempts done on the same dataset by other researchers. The authors made an improvement to their predecessors but not by a substantial margin.

Varshney et al, A.M.2017 [5] The researchers used three classification algorithms and aimed to compare the results to find out which one was more suitable. The dataset collected is a combination of the my personality dataset collected from Facebook along with tweets and data on users gathered from the Twitter API. Their results concluded that the MNB algorithm produced better results than the KNN and SVM algorithms, where it had an accuracy of 60%. however, this result is somewhat dependent on the dataset collected and that shouldn't neutralize the potential of other algorithms that can be used.

Kumar, K.P. and Gavrilova, M.L.2019 [8] The researchers adapted both the big-five personality model along with the Myers & Briggs' 16 Personality Types model and their goal was to score traits on a more complex personality dimension. This paper was one of the more recent ones on the subject and the results they reached was practically applicable due to its high accuracy. The dataset used is The Twitter MBTI Personality Dataset which has large number of users along with their twitter data and personality types, it had 8675 users with 50 tweets each. Their results were an average accuracy across all traits, a Mean square error of 0.235 and an average F1-score of 0.78. However, these results were unbalanced, which affected the average accuracy.

Ergu, İ., Işık, Z. and Yankayış.2020 [12] Another study tried collecting recent data instead of a large dataset and applying the machine learning algorithms on them. This approach was to utilize the automatic personality recognition on turkish tweets, which had no available dataset. They collected data on 51 volunteers and had them provide their personality traits as well as their twitter data. The results weren't consistent and fluctuated between a range of 76% and 96%. Although, this result is not reliable as they were using a dataset of 51 users with 25-50 tweets which is very shallow.

4 What is new in the Proposed Project?

Our project will provide users with a complete analysis report on their personality type from their social media account. This report will include the following:

1. Identifying users' most dominant personality trait.
2. Evaluate the user and generate a detailed report on their personality type. As well as the other traits their personality is leaning towards.
3. Creating a model that will assess the user then fit them with an appropriate career recommendation for their personality.
4. Object detection from images.

5 Proof of concept

Data Pre-Processing:

- Data pre-processing is the method of preparing the raw data and transforming it to an understandable format to enhance the learning of the model. So, we tried to enhance our model by applying different pre-processing techniques which are: data normalization, tokenization, removing the stopwords, Stemming and Lemmatization, and Removing useless strings as hashtags mentions, URLs' and, etc.. . .
- Object detection pre-processing consists of resizing the images in the dataset to fit our model, also we'll use an image labelling or annotation tool to label the images so that they're readable to our model.

6 Project Management and Deliverables

6.1 Deliverables

- Python backend.
- Web application.
- Software Requirement Specification document.
- Software Design document.
- Two conference papers.
- Thesis document.

6.2 Tasks and Time Plan

Here is our time plan showing all stages of the project starting from reading papers, passing by applying pre-processing and feature extraction of data and training the model with machine and deep learning models, until the testing and enhancing stage and the duration of each one is shown in Figure 5.

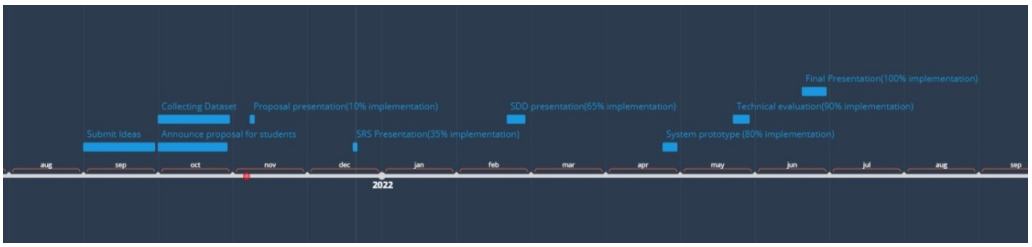


Figure 5: Time Plan.

6.3 Budget and Resource Costs

- Meeting with psychiatrists (300-500EGP Per session)

7 Supportive Documents

7.1 Dataset

The dataset we used contains posts that makes one of the 16 MBTI types as shown in Figure 6 The tags are a combination of four different characters. Each on is corresponding to the one of the trait classes in the four MBTI test. This dataset contains about 8650 rows. The distribution of MBTI traits (number of rows out of 8600) in each class is as follows[13]:

- Introverted (I): 6664; Extroverted(E): 1996
- Sense(S): 7466; Intuition (N): 1194
- Thinker (T): 4685; Feeler (F): 3975
- Judge (J): 5231; Perceive (P): 3429

Each row represents a different user.

Posts From User (Separated by |||)	MBTI type
Newton's Universal Gravity Law. I mean seriously, where would we be if nothing followed that law? Dust particles in space.|||Well, if money and time was no object, I would backpack my way around the world. I would go where ever my feet led me in a westward fashion; only using money when I absolutely needed to. I always...|||http://www.oglaf.com/media/comic/failsafe.jpg|||I'm still laughing!...	INTP

Figure 6: Dataset Sample

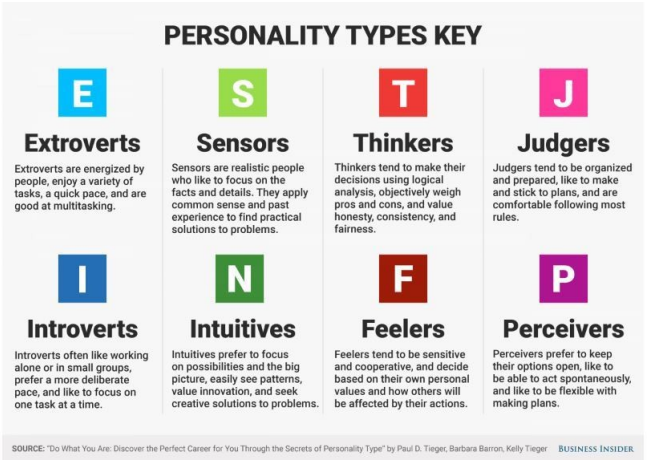
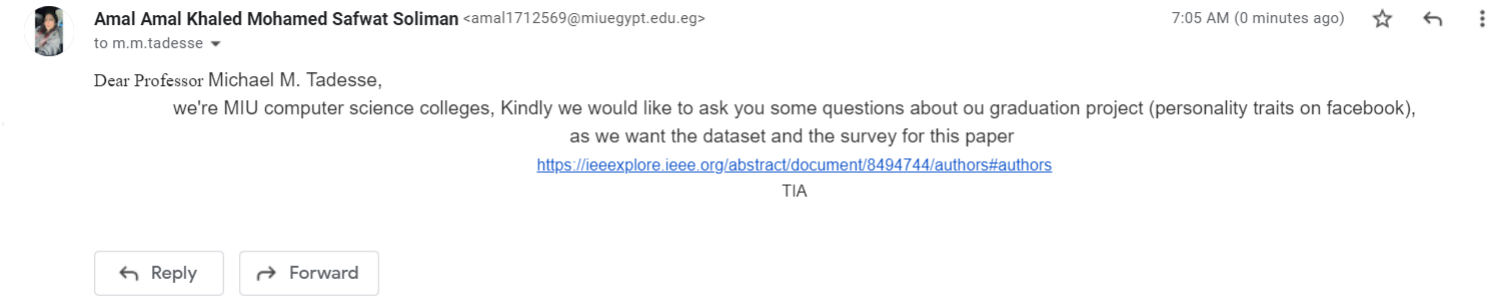


Figure 7: MBTI Types

7.2 Contect Authors

We contacted Professor Michael M. Tadesse to ask for their used dataset to gather a large dataset to train and test our project on.



7.3 Survey

75.% of people answered that they knew about the Personality traits system before in Egypt.While only 24.5% answered that they had never heard about it. This indicates that whenever this system appears it will make a huge buzz.8

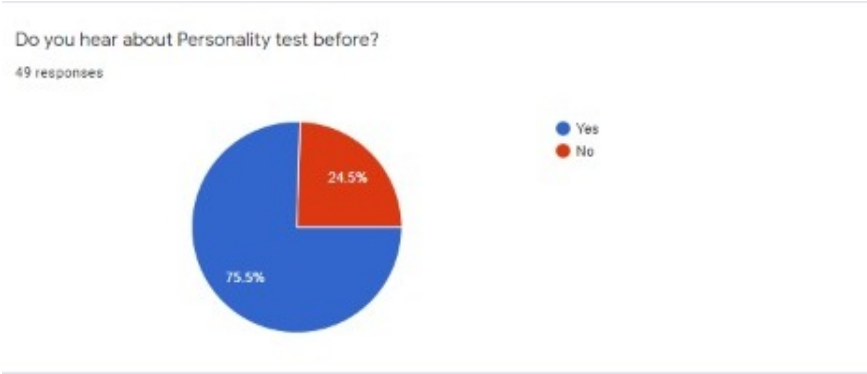


Figure 8: Survey feedback

References

[1] James V Haxby, Elizabeth A Hoffman, and M Ida Gobbini. "The distributed human neural system for face perception". In: *Trends in cognitive sciences* 4.6 (2000), pp. 223–233.

[2] Jennifer Golbeck, Cristina Robles, and Karen Turner. "Predicting personality with social media". In: *CHI'11 extended abstracts on human factors in computing systems*. 2011, pp. 253–262.

[3] Hartayuni Sain and Santi Wulan Purnami. "Combine sampling support vector machine for imbalanced data classification". In: *Procedia Computer Science* 72 (2015), pp. 59–66.

[4] Navonil Majumder et al. "Deep learning-based document modeling for personality detection from text". In: *IEEE Intelligent Systems* 32.2 (2017), pp. 74–79.

[5] Vanshika Varshney et al. “Recognising personality traits using social media”. In: *2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI)*. IEEE. 2017, pp. 2876–2881.

[6] Mariam Hassanein et al. “Predicting personality traits from social media using text semantics”. In: *2018 13th International Conference on Computer Engineering and Systems (ICCES)*. IEEE. 2018, pp. 184–189.

[7] Reinert Yosua Rumagit and A. S. Girsang. *Predicting personality traits of Facebook users using text*. Oct. 2018. URL: https://www.researchgate.net/publication/329031101_Predicting_personality_traits_of_facebook_users_using_text_mining.

[8] KN Pavan Kumar and Marina L Gavrilova. “Personality traits classification on twitter”. In: *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE. 2019, pp. 1–8.

[9] Md Abdur Rahman et al. “Personality detection from text using convolutional neural network”. In: *2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT)*. IEEE. 2019, pp. 1–6.

[10] Marwa S Salem, Sally S Ismail, and Mostafa Aref. “Personality traits for egyptian twitter users dataset”. In: *Proceedings of the 2019 8th International Conference on Software and Information Engineering*. 2019, pp. 206–211.

[11] Hsin-Chang Yang, Chung-Hong Lee, and Chia-Yi Yeh. “Mining Personality Traits from Social Text Messages”. In: *Proceedings of the 7th Multidisciplinary in International Social Networks Conference and The 3rd International Conference on Economics, Management and Technology*. 2020, pp. 1–5.

[12] İzel Ergu, Zerrin Işık, and İsmail Yankayış. “Predicting Personality with Twitter Data and Machine Learning Models”. In: *2019 Innovations in Intelligent Systems and Applications Conference (ASYU)*. IEEE, pp. 1–5.

[13] Siti Khotijah. “Big Five Personality Test + Clustering”. In: (). URL: https://www.kaggle.com/khotijahs1/big-five-personality-test-clustering/data.

[14] Isabel Briggs Myers. *The 16 MBTI® Type*. URL: https://www.myersbriggs.org/my-mbti-personality-type/mbti-basics/the-16-mbti-types.htm.