Predicting Myers-Briggs Personality Types by the Natural Language from Social Media Posts

학번: 2018320252, 이름: Madi Bokishev
지도교수: Hyunwoo Kim

2022.5.27

Abstract

We present a technique for predicting personality types by analyzing posts from social media platforms. People tend to act differently according to their personality traits, which is reflected in the texts they write. Myers-Briggs personality types system divides people into 16 distinct profiles, based on 4 major groups. With the help of this system, in this paper, we aim to predict the personality types by the natural language people used in their social media posts and experimentally unfold the text writing patterns of each 16 personalities. Analyzing the common language traits of each group can serve as a basis for further psychological studies.

1 Introduction

The main concept of the Myers-Briggs Type Indicator (MBTI) personality system is to organize a significantly broad range of people's specific behavior, perceptions, judgments, preferences, and traits variations into a fairly small number of descriptive and understandable personality profiles. The described system distinguishes 16 different personality types and suggests a theory of recognizing individuals' particular traits and patterns of thinking according to their profile type [1]. This theory also implies the common behavior preferences for particular situations of the people with the same MBTI type. This system detailly describes how the unique personality type affects the everyday decision-making propensity of individuals. That is, people with one of the analytical types would prefer to choose a job highly related to logical, intellectually challenging, and technical-oriented processes, while people from the Explorers group tend to dedicate themselves to creative and artistic professions. This system has shown itself as a reliable and accurate tool for categorizing people and finding the most suitable approach for individuals by predicting their behavior patterns and preferences. This method can be used in a significantly large number of areas, from the entertainment industry to recruiting processes. These factors show the importance of distinguishing individuals' personality types for optimizing a broad number of human-related processes. However, examining each person with the corresponding test is a time-consuming process, thus a certain automatization is required. The main aim of this paper is to experimentally conduct an approach for categorizing and predicting MBTI types by analyzing the texts obtained from the most widely used social media platforms, such as Instagram or Twitter. With the help of machine learning algorithms, we will create and train a model that analyzes human language texts, categorizes authors by their writing traits, and predict their personality type.

Contributions. This paper makes the following contributions:

- We present an approach of natural language processing to identify the psychological type of the text authors. The process is automized thus allowing for replacing the manual test passing by individuals.
- MBTI company's website[2] claims a 90% accu-

racy of personal tests. We will determine the accuracy of our model to compare the manual testing approach efficiency with language analysis.

• We will reveal the hidden patterns of writing tendencies of each cognitive group.

2 Overview

This section describes the principles of MBTI and its practical use. Here we will overview the problem and approach of creating an automated way of defying the Myers Briggs Type Indicator and how it can be implemented.

2.1 MBTI Motivation

MBTI or Myers Briggs Type Indicator can be an incredibly useful tool for understanding the underlying patterns of individuals' communication and social interaction patterns. Considering one's type indicator leads to the significant benefits in the adaption of personal approach for a various scope of everyday tasks and situations. The basic principle of MBTI is based on Carl Jung's work and "Psychological Type." The system was created by Isabel Briggs Myers and her mother Katharine Briggs to make Jung's work more approachable and practical in people's daily lives. More than 2 million people use the MBTI personal inventory worldwide, and the theory has been validated by 20 years of research and over 4000 research papers[3]. The most common uses of type indicators are understanding one's preferences, communication and interaction styles, stress reactions, problem-solving approaches, and much more. The goal of MBTI is to help people to become more conscious of their preferences and to consider the benefits of other problem-solving styles. There does not exist a style that would be superior to the others, however, considering all the sides improves adaptation and flexibility.

2.2 MBTI Preferences Scale

Myers Briggs Type Indicator designed and implemented the system of four personal type dimensions, scaling the preferences of people by opposing two cognitive items in each of that dimensions. These scale dimensions are presented in this way:

•**Extraversion - Introversion**: This scale shows where the person tends to direct their energy and put their attention, whether to spend time in the outer world with other people(Extraversion) or in their inner world of ideas and images (Introversion)[4].

•**Sensing - Intuition**: Describes how preferences and tendencies of information perception [5], whether to pay more attention derived through five senses (Sensing) or to pay more attention to the patterns and possibilities from the received information (Intuition).

•**Thinking - Feeling**: Shows the preferences of decision-making, whether it is more based on objective principles and impersonal facts (Thinking) or personal concerns (Feeling)[6].

•**Judging - Perceiving**: The preferences between a more structured and decided lifestyle (Judging) or a more flexible and adaptable lifestyle (Perceiving)[7].

In sake of simplicity the Myers Briggs Type Indicator system uses the abbreviations for each item by defining them with a single letter: **E/I** - Introversion/Extraversion, **S/N** - Sensing/Intuition, **T/F** - Thinking/Feeling, Judging/Perceiving - **J/P**.

The combination of these designations next produces one of the 16 distinct personality types. Each type can be distinguished by its unique preferences, tendencies, and personal traits. Each person who completes the specially designed test is assigned with their personality type. The types are defined using the abbreviation of four letters, derived from the four dimension scales. For instance, the person whose preferences are skewed to Extraversion, Sensing, Feeling, and Judging is assigned with **ESFJ** personality type.

2.3 MBTI Practical Use

The MBTI principles describe the impact of the individuals' preferences on their everyday life and how they interact and cooperate with other people of the same or different personality types. Considering this fact, distinguishing type indicators can have practical application in such areas as careers and personal development. MBTI helps people to understand their natural advantages and disadvantages in various scope of tasks within their working environment. For instance, a lawyer whose personality type is INFP, motivated by internal personal values and new interpretations of the law, tends to act differently from another lawyer of ISTJ type, who is commonly driven by precedents and traditions.

Communication part is one of the most crucial parts of the working environment. Considering the employees' interaction preferences, companies can significantly improve the productivity and efficiency of the processes where personal interactions are required.

Integrating the MBTI system can sufficiently improve the process of team-building of the company's employees. A better way of communication, identifying individuals' similarities and differences, stimulation, and job delegation by considering the MBTI types of each member allows the more efficient team-maintaining and coordination.

Even considering the fact the recruiting process should not be based only on personality type, the integration of MBTI into the existing working environment can greatly improve the overall efficiency and productivity of the company[8].

Also, MBTI analysis can contribute to the psychological studies of human behavior, based on their preferences and tendencies, which will further find its scientific and practical applications in different areas, such as human studies.

For ethical and moral reasons, it should be clear that all psychological tests and experiments must be conducted with the agreement of the person who is a subject of those.

2.4 Automatization the process of MBTI defining

This research paper suggests an alternative to the traditional way of defining the individual's personality type. By applying the machine learning algorithms and sentiment analysis on the social network posts, the expected results should reveal the person's preferences based on their writing style. The main goal of the research is to distinguish the personality types using the limited data of the social networks posts, to define the correlation between each type of indicator and text-writing patterns.

To complete this task, our team designed a machine learning algorithm with implemented natural language processing techniques. Analyzing the human speech from the dataset derived from each distinct cognitive group will create a possibility to analyze the styles and features of each type.

This research contributes to the benefits of automatizing the process of MBTI type testing by post scanning and analyzing, which eliminates the necessaries for manual test passing. Moreover, it provides a new way of analyzing the personality type without direct human interaction, which can potentially eliminate the bias during personal test completion. A remarkable percentage of people tend to not respond to the provided test questions with full honesty which can lead to statistical bias, occasionally making the derived source of data not always credible. The automated approach can lead to sufficient improvement in statistics by neglecting the fact of human nature prejudice. This project also contributes the psychological studies with the scientific value in terms of defining the hidden writing style and cognitive patterns of different individual groups

Further in this paper we will detail describe the problem, the process of gathering the data, prepossessing and filtering the dataset, the main approach, evaluation criteria, and obtained results. According to these results, we will define the directions for further studies.

3 Problem

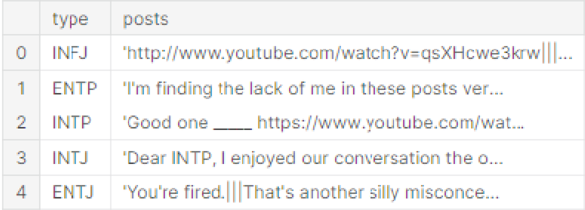
This section describes the used dataset, the process of collecting, prepossessing, visualization, and analyzing the data in terms of natural language.

3.1 Dataset

For this problem, we have used the Kaggle (MBTI) Myers-Briggs Personality Type Dataset. This dataset roughly consists of 8600 rows that contain persons' MBTI type in form of four letter codes (INTJ, ENFP, ESTJ ...), as well as their 50 most recently published posts. This dataset was created by collecting the information through the PersonalityCafe forum. The dataset includes a high variance of people and their MBTI types[9].

3.2 Data collecting and visualization

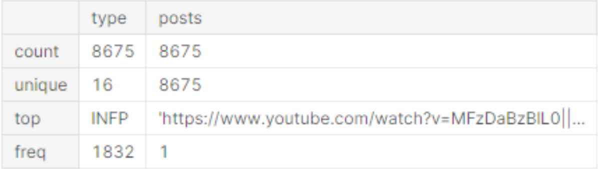
Initially, the derived data is represented in the following way:



	type	posts
0	INFJ	'http://www.youtube.com/watch?v=qsXHcwe3krw|||...
1	ENTP	'I'm finding the lack of me in these posts ver...
2	INTP	'Good one _____ https://www.youtube.com/wat...
3	INTJ	'Dear INTP, I enjoyed our conversation the o...
4	ENTJ	'You're fired.|||That's another silly misconce...

Figure 1: Five first rows of dataset

In Figure 1, the left column of the table shows the MBTI code, while the right column contains all correspondingly published posts. Exploring the dataset, we obtain a global vision of frequencies and types. 16 types of unique personality indicators, with 1832 occurrences are detected. No repeated posts are observed.

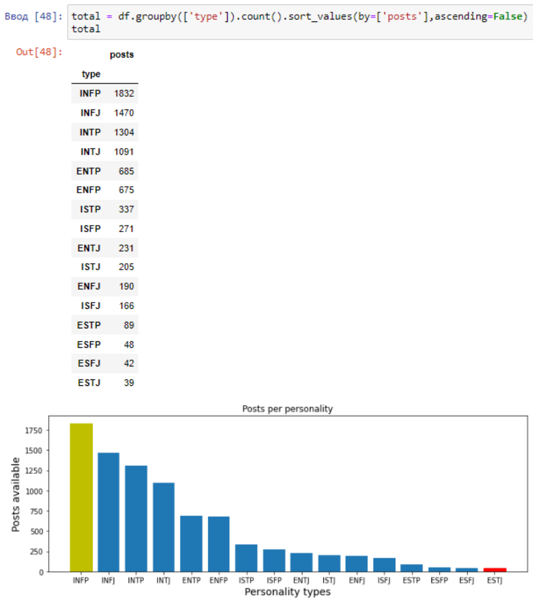


	type	posts
count	8675	8675
unique	16	8675
top	INFP	'https://www.youtube.com/watch?v=MFzDaBzBlL0||...
freq	1832	1

Figure 2: Data Frequency

The next figure demonstrates the number of posts grouped by corresponding personality types. By applying this method, it is observed that the given data is unbalanced in the sense that there is no equal distribution for the number of posts per personality, as can be seen in the following graph (Figure 3)

By continuing the data exploration, we show the obtained histogram of the data distribution observed for each column (Figure 4). The resulting graph concludes that most posts are between 7,000 and 9,000 words long.



```
Ввод [48]: total = df.groupby(['type']).count().sort_values(by=['posts'],ascending=False)
           total
```

```
Out[48]:
              posts
        type
        INFP   1832
        INFJ   1470
        INTP   1304
        INTJ   1091
        ENTP    685
        ENFP    675
        ISTP    337
        ISFP    271
        ENTJ    231
        ISTJ    205
        ENFJ    190
        ISFJ    166
        ESTP     89
        ESFP     48
        ESFJ     42
        ESTJ     39
```

Figure 3: Unbalanced distribution of data

The shown line represents the kernel density estimation. It is a fundamental data smoothing problem where inferences are made about the population, based on a finite sample of data. This kernel density estimation is a function defined as the sum of kernel functions at each data point.

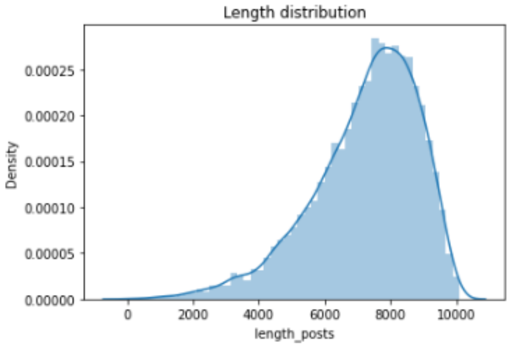
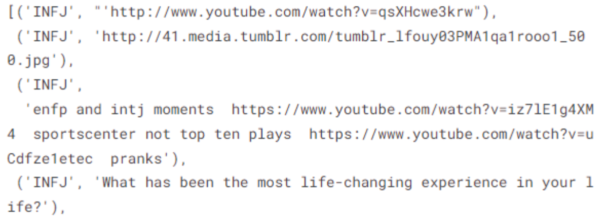


Figure 4: Posts' lengths distribution

3.3 Data preprocessing

During the detailed data analysis, it was defined that the dataset contained a significant amount of redundant information, such as hyperlinks (Figure 5). Since we assume that hyperlinks do not provide any meaningful information about users' personalities, moreover might cause a statistical bias, they are ignored during the testing and dropped in the preprocessing stages.



```
[('INFJ', "'http://www.youtube.com/watch?v=qsXHcwe3krw"),
 ('INFJ', 'http://41.media.tumblr.com/tumblr_lfouy03PMA1qa1rooo1_50
0.jpg'),
 ('INFJ',
  'enfp and intj moments  https://www.youtube.com/watch?v=iz7lE1g4XM
4  sportscenter not top ten plays  https://www.youtube.com/watch?v=u
Cdfze1etec  pranks'),
 ('INFJ', 'What has been the most life-changing experience in your l
ife?'),
```

Figure 5: Redundant information

Data cleaning process also included such actions as removing all punctuation characters and normalizing the words into lower case. An important part of this stage is filtering the stop-words. Despite being the most common words in the English language, stop words, such as "the", "is", "a", "an" etc, are not considered to be valuable information in terms of natural language processing[10]. Such words are removed from the dataset to acquire a higher accuracy during training and testing steps.

For a clearer picture of the updated dataset content, the wordclouds in figures 6-8 represent the most frequently occurred words within the whole dataset, and for particular personality types respectively.

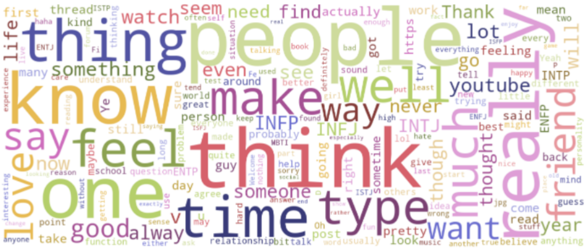


Figure 6: Most common words in dataset

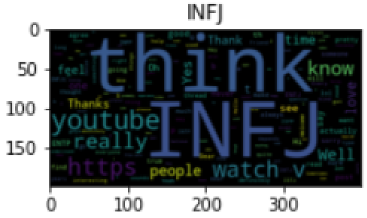


Figure 7: Most common words for INFJ

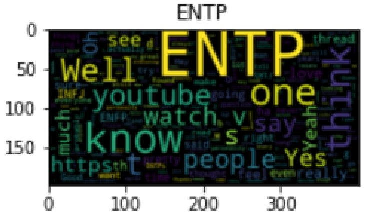


Figure 8: Most common words for ENTP

The final step in data preprocessing is splitting the obtained data into training and testing datasets, with a ratio of 75% to 25%, respectively.

4 Approach

This section demonstrates the machine learning approach to solving the predicting problem using several different models. Here we describe the process of model selection, training, validation, and testing.

4.1 Models selection

For this problem, three different separate models were implemented. After thorough consideration and research held, our team decided on using the following models[11-13]:

- Multinomial Logistic Regression
- Linear Support Vector classifier
- XGBoost Classifier

4.2 Multinomial Logistic Regression

Based on numerous independent factors, multinomial logistic regression is used to predict categorical placement in or the probability of category membership on a dependent variable. The dependent variables might be dichotomous or continuous. Multinomial logistic regression is a straightforward extension of binary logistic regression that allows for the inclusion of more than two categories of the dependent or outcome variable. Multinomial logistic regression, like binary logistic regression, evaluates the probability of category membership using maximum likelihood estimation[13].

After integrating and training our model, the obtained accuracy achieved 72%, while test accuracy was 63%. A sufficient difference between training and testing results was observed, which may address the over-fitting problem.

4.3 Linear Support Vector classifier

Support Vector Classifiers are non-linear, non-parametric techniques used for both classification and regression problems. It has shown promising results in a large scope of fields. SVM is a supervised algorithm that classifies or separates data using hyperplanes. Due to its simplicity yet efficiency, this technique frequently finds its place in various natural language processing tasks.

After training this model on our dataset, we obtained the following results:

Training accuracy: 82%

Testing accuracy: 66%

Even though the model demonstrated significantly better results during training, and slightly higher accuracy on the test dataset than logistic regression, the high difference between test and training performance points to the same overfitting problem.

4.4 XGBoost Classifier

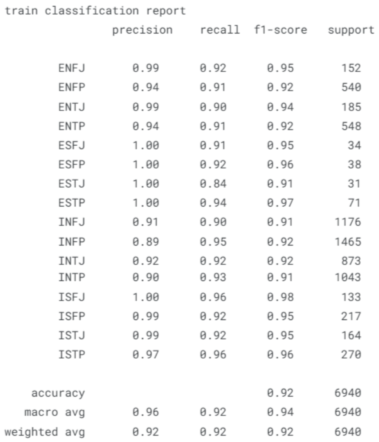
XGBoost or Extreme Gradient Boosting is a scalable, distributed gradient-boosted decision tree machine learning technique, which is commonly used for classification, regression, and ranking problems. Compared to other algorithms, XGBoost and XGBoost machine learning models have a great balance of prediction performance and processing time, which makes them an excellent solution for a large number of problems in NLP fields.

Fitting the model based on the XGBoost algorithm led to the following results.

Training accuracy: 92%

Testing accuracy: 67%

The detailed report of both training and testing processes is shown in the figures below:

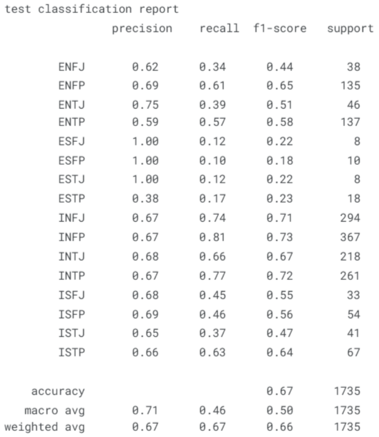


```
train classification report
              precision    recall  f1-score   support

        ENFJ       0.99      0.92      0.95       152
        ENFP       0.94      0.91      0.92       540
        ENTJ       0.99      0.90      0.94       185
        ENTP       0.94      0.91      0.92       548
        ESFJ       1.00      0.91      0.95        34
        ESFP       1.00      0.92      0.96        38
        ESTJ       1.00      0.84      0.91        31
        ESTP       1.00      0.94      0.97        71
        INFJ       0.91      0.90      0.91      1176
        INFP       0.89      0.95      0.92      1465
        INTJ       0.92      0.92      0.92       873
        INTP       0.90      0.93      0.91      1043
        ISFJ       1.00      0.96      0.98       133
        ISFP       0.99      0.92      0.95       217
        ISTJ       0.99      0.92      0.95       164
        ISTP       0.97      0.96      0.96       270

    accuracy                           0.92      6940
   macro avg       0.96      0.92      0.94      6940
weighted avg       0.92      0.92      0.92      6940
```

Figure 9: XGBoost training



```
test classification report
              precision    recall  f1-score   support

        ENFJ       0.62      0.34      0.44        38
        ENFP       0.69      0.61      0.65       135
        ENTJ       0.75      0.39      0.51        46
        ENTP       0.59      0.57      0.58       137
        ESFJ       1.00      0.12      0.22         8
        ESFP       1.00      0.10      0.18        10
        ESTJ       1.00      0.12      0.22         8
        ESTP       0.38      0.17      0.23        18
        INFJ       0.67      0.74      0.71       294
        INFP       0.67      0.81      0.73       367
        INTJ       0.68      0.66      0.67       218
        INTP       0.67      0.77      0.72       261
        ISFJ       0.68      0.45      0.55        33
        ISFP       0.69      0.46      0.56        54
        ISTJ       0.65      0.37      0.47        41
        ISTP       0.66      0.63      0.64        67

    accuracy                           0.67      1735
   macro avg       0.71      0.46      0.50      1735
weighted avg       0.67      0.67      0.66      1735
```

Figure 10: XGBoost testing

5 Evaluation

This section describes the obtained results, analysis, limitations, and considerations of the experiments. We compare the performance of the trained models, we also demonstrate the results of the algorithm in practice examples.

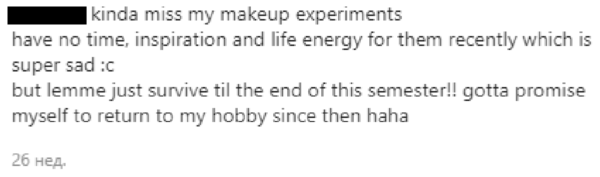
5.1 Results

After comparing the implemented models, it became clear that XGBoost Classifier shows the best values-predicting performance in terms of accuracy for both training and testing on the provided dataset. The final results showed 67% test accuracy, which is significantly lower compared to one of training, which conducted 92% accuracy of personality type predicting.

Even though Linear Support Vector Classifier outperformed Multinomial Logistic Regression during model fitting, on the testing step they showed relatively similar results in terms of accuracy.

5.2 Practice example

For this example, we use XGBoost Classifier to determine the personality type of the author by the provided posts of them. Our trained model analyzed 10 different posts from the provided Instagram page. We include one example of them in this paper. The output results are presented on the pie chart below



[redacted] kinda miss my makeup experiments
have no time, inspiration and life energy for them recently which is super sad :c
but lemme just survive til the end of this semester!! gotta promise myself to return to my hobby since then haha

26 нед.

Figure 11: Post Example

The obtained results determine the author's personality type as INTP, which matches the results of the test provided on the official website of the MBTI company.

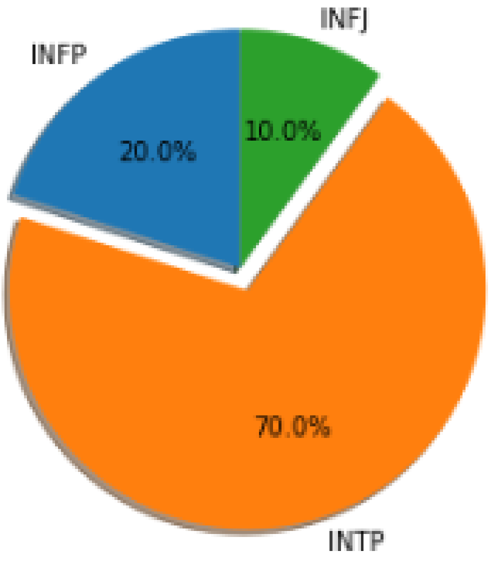


Figure 12: Analysis results

5.3 Limitations and Considerations

By comparing all the obtained model results, the same problem of the high difference between training and testing accuracy is observed for each of the implemented models.

The great performance while testing and relatively low accuracy results on the test set point to the problem of model overfitting. One of the possisolutionstion would be increasing the size of the dataset, fitting the model with more values of representative data can potentially improve the overall performance of the trained model on the test set[14]. During the training process it appeared that additional hyper-parameters of tuning of models did not lead to sufficiently better results on the test set, which indicates the importance of the overfitting issue. Enlarging the dataset should be considered the main approach for further studies.

These limitations of the implemented algorithm become a significant obstacle restricting the practical usage of the proposed method.

6 Conclusion

In this paper, our team proposed a method of automatically analyzing social media posts with aim of recognizing the distinct personality types of the users. We implemented and presented a machine learning approach to solving the predicting problem for defining Myers-Briggs personality types indicators by processing the natural language of written form. By fitting thdifferentrent models, we found XGBoost Classifiertos one of the most suitable decision for accomplishing the given task.

Despite not remarkably high test accuracy, the obtained results indicate the presence of a correlation between the types of users' personalities suggested by the MBTI system and the natural language they tend to use while expressing their thoughts in the written form. Results also demonstrate the existence of writing patterns in distinct cognitive groups. Further research on a larger dataset may lead to clearer results and higher performance in terms of personality types recognition accuracy.

References

[1] Geyer, Peter & Jung, C. (2013). Fundamentals of Personality Type 3. C.G.Jung and Personality Type. 10.13140/2.1.2923.2009. Retrieved from https://www.researchgate.net/publication/264789613_Fundamentals_of_Personality_Type_3_CGJung_and_Personality_Type

[2] The Myers & Briggs Foundation - Reliability and Validity. (2022). Retrieved from https://www.myersbriggs.org/my-mbti-personality-type/mbti-basics/reliability-and-validity.htm#:~:text=Facts%20about%20the%20MBTI%C2%AE,to%2090%25%20of%20the%20time.

[3] Hargreaves C. 2022. MBTI[PowerPoint slides]. Professional Development Programme for Master's Students. Imperial Colledge London. https://www.imperial.ac.uk/media/imperial-college/administration-and-support-services/staff-development/public/impex/MBTI.pdf

[4] The Myers amp; Briggs Foundation - extraversion or introversion. (n.d.). Retrieved from https://www.myersbriggs.org/my-mbti-personality-type/mbti-basics/extraversion-or-introversion.htm

[5] The Myers amp; Briggs Foundation - sensing or intuition. (n.d.). Retrieved from https://www.myersbriggs.org/my-mbti-personality-type/mbti-basics/sensing-or-intuition.htm

[6] The Myers Briggs Foundation - Thinking or Feeling. (2022). Retrieved from https://www.myersbriggs.org/my-mbti-personality-type/mbti-basics/thinking-or-feeling.htm

[7] The Myers Briggs Foundation - Judging or Perceiving. (2022). Retrieved from https://www.myersbriggs.org/my-mbti-personality-type/mbti-basics/judging-or-perceiving.htm

[8] Applications of the MBTI — By Peter Geyer. (2022). Retrieved from https://www.personalitypathways.com/MBTI_geyer-2.html#:~:text=The%20MBTI%20has%20applications%20in,change%20situations%20in%20different%20ways.

[9] (MBTI) Myers-Briggs Personality Type Dataset. (2022). Retrieved from https://www.kaggle.com/datasets/datasnaek/mbti-type

[10] Khanna, C. (2021). Text pre-processing: Stop word removal using different libraries. Retrieved from https://towardsdatascience.com/text-pre-processing-stop-words-removal-using-different-libraries-f20bac19929a

[11] What is XGBoost?. (2022). Retrieved from https://www.nvidia.com/en-us/glossary/data-science/xgboost/

[12] Auria, L., Moro A. R. (2008). Support Vector Machines (SVM) as a technique for solvency analysis [Ebook]. Retrieved from https://www.econstor.eu/bitstream/10419/27334/1/576821438.PDF

[13] Starkweather, J., Kay Moske, A. (2011). Multinomial Logistic Regression [Ebook]. Retrieved from https://it.unt.edu/sites/default/files/mlr_jds_aug2011.pdf

[14] Ying, Xue. (2019). An Overview of Overfitting and its Solutions. Journal of Physics: Conference Series. 1168. 022022. 10.1088/1742-6596/1168/2/022022. Retrieved from https://www.researchgate.net/publication/331677125_An_Overview_of_Overfitting_and_its_Solutions