**Personality Prediction using MBTI**

A report submitted in partial fulfillment of the

Requirements Of

Mini-Project(IS65)

6th Semester By

**1MS18IS115 Tanishq Soni**

**1MS18IS133 Lakshya Aditi Sinha**

**1MS18IS086 Sainya Goyal**

**1MS18IS078 Rithik Singhai**

Under the guidance of

**Prof. Lincy Mathews**

Dept. of ISE, RIT

****

DEPARTMENT OF INFORMATION SCIENCE & ENGINEERING

RAMAIAH INSTITUTE OF TECHNOLOGY

(AUTONOMOUS INSTITUTE AFFILIATED TO VTU)

M. S. RAMAIAH NAGAR, M. S. R. I. T. POST, BANGALORE – 560054

2020-2021

**Acknowledgement**

We are overwhelmed in all humbleness and gratefulness to acknowledge our depth to all those who have helped us to put forth these ideas, well above the level of simplicity and into something concrete.

We acknowledge the guidance and support extended by our guide, *Prof. Lincy Mathews*, Department of ISE. Her incessant encouragement and invaluable technical support has been of immense help. Her guidance gave us the environment to enhance our knowledge and skills. We also take this opportunity to thank our Principal and college management for providing us with the platform to perform. We learnt many new things along the way. We are really thankful to them.

## Abstract

The MBTI model was designed to assess psychological preferences , how they interact with the world and how they make decisions.

Personalities can be assessed via inventories, otherwise known as questionnaires or scales of some kind. A respondent may be given a list of statements in a standard order with a specific set of answer choices. The respondent writes down the degree to which they agree or disagree with a statement on a scale from strongly disagree to strongly agree or whether they agree or disagree in general by answering true or false. The assessment of a person’s personality is important and helpful for managing people and for understanding yourself. It also provides insight into the human psyche.

Our project aims to predict a person’s personality using various Machine-Learning, and Neural Network classification models. Furthermore, we aim to compare the performance of different models and find the best performing model. We use the MBTI (Myers-Briggs Type Indicator) dataset, which is a self-inventory designed to identify a person's personality type, strengths, and preferences. In MBTI, People are identified as having one of 16 personality types (Introversion vs Extroversion, Sensing vs Intuition, Feeling vs Thinking, Judging vs Perceiving, Interpreting your results) which will be our attributes on which we a person’s personality would be assessed. We intend to use Natural Language Processing techniques for pre-processing the MBTI dataset to create a bag of words which would be used to train and test the models.

**Contents**

[1 Introduction](#_heading=h.30j0zll)

[1.1 Motivation](#_heading=h.1fob9te)

[1.2 Scope](#_heading=h.3znysh7)

[1.3 Objectives](#_heading=h.2et92p0)

[1.4 Proposed Model](#_heading=h.tyjcwt)

[1.5 Organisation of Report](#_heading=h.3dy6vkm)

[2 Literature Review](#_heading=h.1t3h5sf)

[3 System Analysis and Design](#_heading=h.4d34og8)

[4 Modelling and Implementation](#_heading=h.2s8eyo1)

[5 Testing & Results](#_heading=h.17dp8vu)

[6 Conclusion and Future Work](#_heading=h.3rdcrjn)

**Chapter 1**

# Introduction

Based on the answers to the questions on the inventory, people are identified as having one of 16 personality types. The goal of the MBTI is to allow respondents to further explore and understand their own personalities including their likes, dislikes, strengths, weaknesses, possible career preferences, and compatibility with other people.

No one personality type is "best" or "better" than another. It isn't a tool designed to look for dysfunction or abnormality. Instead, its goal is simply to help you learn more about yourself. The questionnaire itself is made up of four different scales.

## Motivation

Experts agree that one of the most common reasons for people to get the wrong Type when they take a personality test is because they answer based on not how they really are, but how they want to be or how they think they should be.

People may not have a clear picture of themselves. Your 4 letter Personality Type stays the same from birth to death, but it's easier to determine one's Type when one is under 25. The reason is that as one matures and becomes more well rounded, one learns to adapt to situations. You pick up learned behaviors and these become layers of personality on top of your original personality. Thus, the older and more mature you are, the more layers you have hiding your original personality.When you are trying to identify your 4 letter Personality Type, what you want to identify is what is known as your "True Self," or your "Core Self." This would be your personality without all the learned behaviors.

The learned behaviors are like a layer of personality to your Core Self. This layer is known as your "Developed Self."There is one additional layer and that is your "Contextual Self" which describes how you behave in different circumstances. In certain situations, you might act totally differently than what your Core Self would suggest.

Personality prediction is important as there’s significant correlation between personality and real-world behavior. When we focus on computational recognition of personality, this technique is more and more promising.In the business field,personalized marketing and application design can help companies make great success in gaining more customers.

* 1. **Scope**

We set out to predict one’s MBTI personality type from

a test we contact through a series of questionnaires. Our algorithm takes in an excerpt of text as input and outputs the predicted MBTI personality label (e.g. ENTJ). We will survey a variety of methods for this task, looking both at classical Supervised Learning and at the efficacy of deep learning with actively trained word embeddings on such a task. Then we perform comparisons and analysis on their resulting error and accuracy to find the method that is most effective for this problem.

## Objectives

Knowing your 4 letter Personality Type is extremely powerful when it comes to understanding yourself and others. There are hundreds of books written about Personality Type that you can access to boost your personal growth and your success.

If you have the wrong Personality Type, all of this information is useless.

Thus if you want any of the following, then making sure you have the correct 4 Letter Personality Type is critical:

* You are looking for insight into which careers you might be good at
* You would like insight into your strengths and weaknesses
* You want to achieve more and be more successful
* You want to be happier
* You want to know how to sell to others better
* You want to know how to be more persuasive
* You want to have better relationships
* You want to know how to get along with others better
* You have trouble getting along with a person and you want help figuring out what to do

The Personality Type system will help you to achieve all of those things, but only if you know your Type. Otherwise, all of the information will be inaccurate.



## Proposed Model

The implementation of the personality prediction model consists of:

* User Interface : The user answers a set of 20 questions. These questions are exclusively determined in a way to determine one’s personality, ie. which one of the 16 personality types he/she belongs to. The answers to these questions are stored in the form of a text file.
* Python Model:It pre-processes the responses obtained from the user and uses various classification algorithms to predict the different personality traits. Finally, the personality of the user is obtained.
* NLTK library was used to classify the personality type. But implementing the outcome into 16 different categories resulted in a very bad probability of the model with only 43% accuracy on the train dataset and 12% accuracy on the test dataset. Hence, 4 different classifiers were created based on the 4 MBTI traits. This improved the accuracy of the results.
* The model uses an ensemble learning algorithm to detect the personality of the user. The ensemble learning combines the following 4 algorithms: Naive Bayes Classifier, Multinomial Naive Bayes Classifier, and Bernoulli Naive Bayes Classifier. The final accuracy of the model is:
* The output received in the above model is then displayed to the user in the form of a four letter Personality Type.

**1.5 Organisation of Report**

In order to explain the developed system, the following sections are covered:

● **Literature Review** describes the study of the existing systems and techniques taken into account prior to development of the proposed system.

● **System Analysis and Design** provides a detailed walk through of the software engineering methodology adopted to implement the model, an overview of the system and the modules incorporated into the system 4

● **Modelling and Implementation** provides a deeper insight into the working of the model. The various modules and their interactions are depicted using relevant descriptive diagrams.

● **Testing** the model to ensure bug/error free model along with the **Results** obtained. Discussion then provides detailed analysis on quality assurance measures.

● **Conclusion** about the Results obtained after successfully running the model and **Future Scope** of the model is highlighted.

**Chapter 2**

# Literature Review

In paper [1], the researchers built an architecture for the classification of various personality attributes derived from Twitter posts which consist of a mixture of word embeddings and number-based encoding along with linguistic features. The MBTI dataset was used for validating and training the machine learning algorithms. SVM and XGBoost were implemented for the classification of different personality traits. After preprocessing, the GloVe word embedding and TF-IDF vectorizer were used to create a weighted aggregation of vector representation of every related text document, which was then fed to the classification algorithm to predict the appropriate personality type.

The authors of the paper [2] suggested a model based on latent features to predict a person’s personality traits. The PANCLEF dataset was used for experimentation and evaluation. In the initial step, every text document is divided into five attributes drawn on matters such as URLs, mentions, phrases, emojis, and unimportant terms such as stopwords, short or long words, punctuation are removed from the text document. With the help of TF-IDF, the preprocessed texts are translated into a matrix representation. Further, LDA, NMF, and PCA selection models were used to extract the latent features via matrix representation. Random forest (RF), logistic regression (LR), and support vector classifier (SVC) supervised algorithms were used for classification.

The researchers of this paper [3] used the open-source MBTI (Merys-Briggs Type Indicator) dataset and proposed a framework that includes four steps. In the initial step, text pre-processing has been performed with the help of lemmatization, similarity hashing, and named entity recognition to remove the irrelevant features from the input document. Next, a feature word vector comprising significant features was built by using the word2vec embedding technique and TF-IDF vectorizer. In the third step, the training and validation of the extracted features were performed by using supervised classification algorithms- random forest and XGBoost. The performances of the trained models were evaluated in the final step based on f1-score, recall, and accuracy. Afterward, unseen or new test written documents were utilized to train the models.

In the paper[4], the myPersonality dataset provided by Facebook was used to detect the personality of their users. Initially, OpenNLP was employed to preprocess the input data. Next, feature extraction was performed with the help of Social Network Analysis (SNA), Structured Programming for Linguistic Cue Extraction(SPLICE), and Linguistic Inquiry and Word Count(LIWC). The Pearson correlation analysis was used for selecting the features. The extracted features were fed to four different machine learning models- SVM, Gradient Boosting, Logistic Regression, and XGBoost. The personality prediction was done based on Big Five personality traits:

Openness, conscientiousness, extraversion, agreeableness, and neuroticism. On the whole, the average accuracy of the model is 74.2%.

In this paper [5], three-modal architecture was proposed with a separate channel to anticipate the Big Five Personality attributes with the use of text, audio, and video data. The ChaLearn First Impressions Dataset and the ChaLearn Validation Set consisting of YouTube vlogs were utilized as the training set and the test data for the model respectively. There are three different channels each for audio, video, and text data for which Convolutional Neural Networks are implemented. The channels are merged by integrating their fully connected layers and on the decision level. It was proved that each modality comprises a particular signal indicating the personality prediction, but combining the three modalities outperformed the individual ones.

The researchers of the paper [6] suggested the design for the analysis of text from Social Media and hence anticipate the personality of brands. The personality of the brand was predicted based on the Big Five personality traits- Openness, Consciousness, Extroversion, Agreeableness, and Neuroticism. This result was a great boon to the brand as it helped them to modify their Marketing Strategies and at the same time develop a better understanding of their Clients. The individual brand page data-set along with the MyPersonality data-set was used to extract the important features. Two approaches known as Pearson Correlation and Gradient Boosting were employed to do feature selection on various machine learning algorithms such as Gradient Boosting, Feed- Forward Neural Network, and Support Vector Regression (SVR). From the above three algorithms, the accuracy of XGB was the highest for personality prediction.

Di Xue et. al. [7] believed that language is a better mode of expressing one’s feelings and thoughts for other people to comprehend, therefore text can be effectively used to predict one’s personality traits. The AttRCNN model was suggested which is a two-layer deep learning framework, based on RCNN for vectorization of sentences to abstract the semantic vectors. The Big Five design was implemented on the MyPersonality dataset consisting of information of millions of Facebook users. The machine learning algorithms used in this model were Random forest classifier, SVR, and Gradient Boosting.

**Chapter 3**

# System Analysis and Design

## System Analysis

The Machine learning models we have used to predict personality are as follows :

Naive bayes classifier

Naive Bayes classifiers are a collection of classification algorithms based on Bayes’ Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other.

We used Naive Bayes classifier because it is a fast and straightforward algorithm, and performs well on large chunks of data in case of text classifiers.

Bernoulli Naive bayes

Bernoulli Naive bayes provide more efficient performance though. Faster calculation times come from restricting the data to a numeric 0-1 matrix and taking advantage of linear algebra operations. Sparse matrices of class "dgCMatrix" (Matrix package) are supported in order to furthermore speed up calculation times.

Since Bernoulli Naive Bayes works well for discrete data, where features are only in binary form we used in our 4 four binary classifiers to classify each attribute of personality.

Multinomial Naive bayes

Multimodal naive bayes (also known as Multinomial Naive Bayes) is a specialized version of naive bayes designed to handle text documents using word counts as it's underlying method of calculating probability.

## Understand the working of Machine learning Model

In the MBTI dataset, two dimensions I/E and S/N are highly imbalanced, therefore a class balance technique is used for better prediction performance. While training the dataset on different supervised machine learning algorithms, the following outcomes were observed. KNN and SVM classifiers gave overall low performance. The outcome of the Decision Tree algorithm for I/E and S/N traits is better than F/T and J/P traits. Random Forest gave highest for all traits. However, for J/P lowest Recall was obtained using Random Forest. On the other hand, the results obtained by Naives Bayes Classifier were comparatively better. Similarly, using Multinomial and Bernoulli Naive Bayes increased the accuracy of the model. Thus, an ensemble algorithm consisting of the above three algorithms was used to train the MBTI dataset and obtain the required result, i.e. personality of the user.

## 

## System Design

* We have created a set of 20 questions. These questions are exclusively determined in a way to determine one’s personality, ie. which one of the 16 personality types he/she belongs to.The answers to these questions are stored in the form of a text file.NLTK library was used to classify the personality type. But implementing the outcome into 16 different categories resulted in a very bad probability of the model with only 43% accuracy on the train dataset and 10-14% accuracy on the test dataset.Hence, 4 different classifiers were created based on the 4 MBTI traits. This improved the accuracy of the results to 60-61% for every MBTI trait.
* The model uses an ensemble learning algorithm to detect the personality of the user. The ensemble learning combines the following 4 algorithms: Naive Bayes Classifier, Multinomial Naive Bayes Classifier, Bernoulli Naive Bayes Classifier and Decision Tree Classifier.
* The final accuracy of the model is: Between 58-61
* The output received is then displayed to the user.

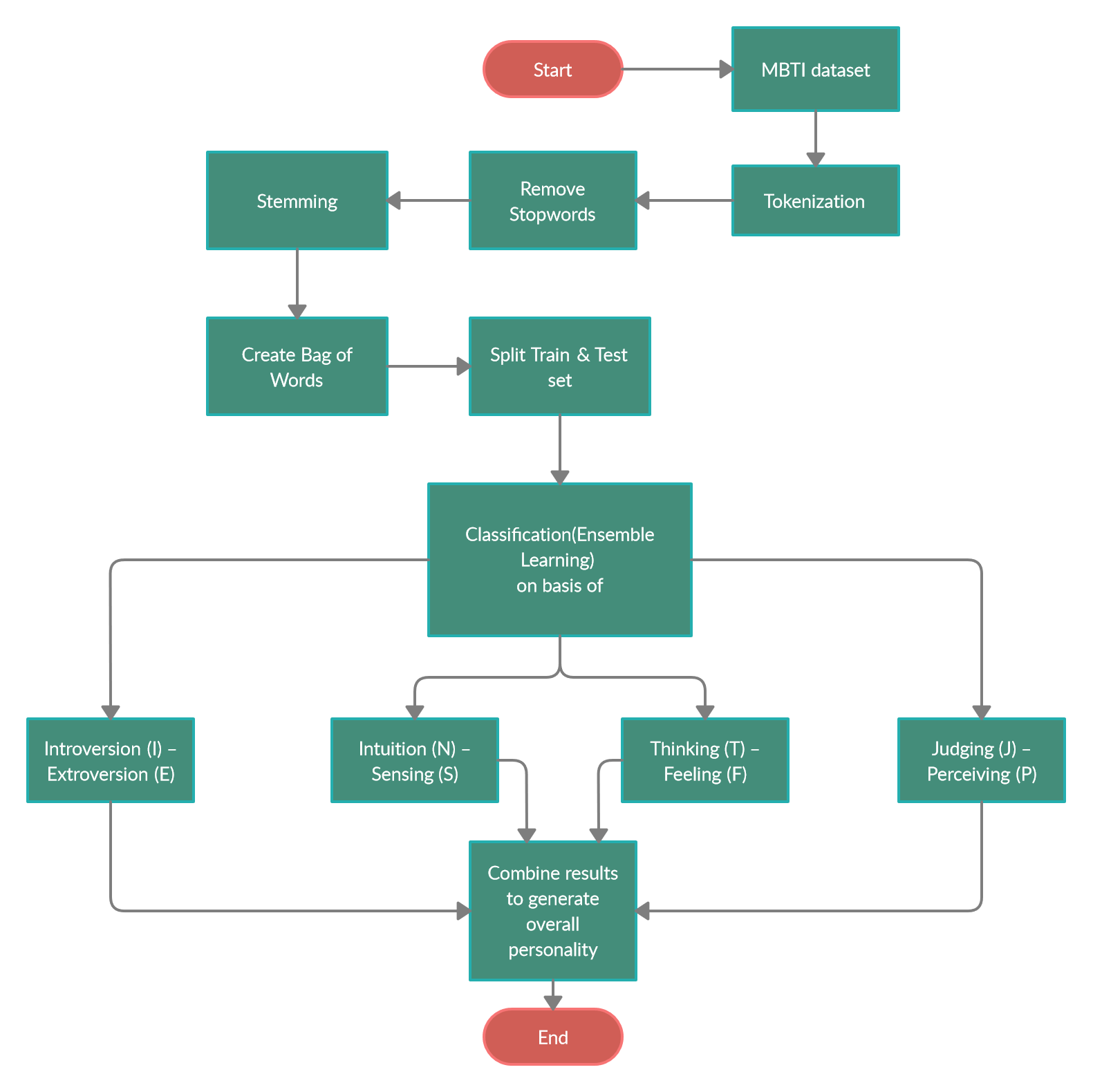
**Chapter 4**

Modelling and Implementation

## 4.1 Use Case Diagram



## 4.2 Workflow Diagram



**Chapter 5**

**Testing & Results**

## 5.1 Testing

From the bag of words we observed that certain keywords were more frequent in certain personalities than others, so we decided to test our model on those keywords.

|  |  |  |
| --- | --- | --- |
| **KEYWORD** | **EXPECTED PERSONALITY** | **OUTPUT PERSONALITY** |
| actually | Introvert | Introvert |
| anything | Extrovert | Extrovert |
| never | Intuition | Intuition |
| though | Sensing | Sensing |
| always | Feeling | Feeling |
| new | Thinking | Thinking |
| usually | Perceiving | Perceiving |
| definitely | Judging | Judging |

## 5.2 Results

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **I VS E** | **N VS S** | **T VS F** | **J VS P** |
| **NaiveBayes** | 58.11 | 54.43 | 59.47 | 54.37 |
| **MultinomialNB** | 60.79 | 56.74 | 60.48 | 54.50 |
| **BernoulliNB** | 61.05 | 57.08 | 60.60 | 54.51 |

**Chapter 6**

# Conclusion and Future Work

**Conclusion**

Personality detection model is capable of classifying a user’s personality into 16 different personalities using 4 binary classifications with accuracy just above 60%. For this we used 20 question responses from the user with a hard voting (ensemble learning) approach on Naive bayes, Binomial Naive Bayes, and Multinomial Naive Bayes classifiers.

Hence, 4 different classifiers were created based on the 4 MBTI traits. This improved the accuracy of the results to 60-61% for every MBTI trait.

**Future Work**

We propose to use this MBTI model and emotion detection methods to extract emotion from text input. Both the keywords and Affect Bearing Word (ABW) will be the best topic of our project to observe feelings from text. Experiments well-tried that human motion was deeply trusted in the open-class word of the sentence. As we know, it's still tough to try to do linguistics parsing with machine learning techniques. all the same, some of the linguistics info and emotional keywords like emphatic keywords & direct emotional keywords are total within the system.

Furthermore we can use these machine learning models that would hold some benefit in the field of Mental health. Detecting emotions and predicting personality of a person would be very beneficial while diagnosing a mental illness.

# Bibliography

[1] K. N. P. Kumar and M. L. Gavrilova, "Personality Traits Classification on Twitter," 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 1-8, 2019

[2] D. R. Jaimes Moreno, J. Carlos Gomez, D. Almanza-Ojeda, and M. Ibarra-Manzano, "Prediction of Personality Traits in Twitter Users with Latent Features," 2019 International Conference on Electronics, Communications, and Computers (CONIELECOMP), pp. 176-181, 2019.

[3]”Personality Identification Based On MBTI Dimensions Using Natural Language Processing” Kishan Das, Himanshu Prajapati, Department of Computer Engineering, Silver Oak College Of Engineering and Technology, Ahmedabad, India

[4] M. M. Tadesse, H. Lin, B. Xu and L. Yang, "Personality predictions based on user behavior on the facebook social media platform", *IEEE Access*, vol. 6, pp. 61959-61969, 2018.

[5] Kampman O, Barezi EJ, Bertero D, Fung P (2018) Investigating audio, video, and text fusion methods for end-to-end automatic personality prediction. In: Proceedings of the 56th annual meeting of the association for computational linguistics (volume 2: short papers), vol 2, pp 606–611

[6]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

[7] D. Xue, L. Wu, Z. Hong, S. Guo, L. Gao, Z. Wu, X. Zhong, and J. Sun, “Deep learning-based personality recognition from text posts of online social networks,” Applied Intelligence, vol. 48, no. 11, pp. 4232–4246, May 2018