PERSONALITY PREDICTION FROM TWITTER DATASET USING MACHINE LEARNING

BY

MD. THOUFIQ ZUMMA
ID: 181-15-10968

JERIN AKTHER MUNIA
ID: 181-15-11136

DIPANKAR HALDER
ID: 181-15-11137

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Md. Sadekur Rahman
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised By

Md. Tarek Habib
Assistant Professor
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

04, JANUARY 2022

APPROVAL

This Project/internship titled **Personality Prediction from Twitter Dataset using Machine Learning**, submitted by Md.Thoufiq Zumma, Jerin Akther Munia, Dipankar Halder, ID No: 181-15-10968, 181-15-11136, 181-15-11137 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 04-01-2022.

BOARD OF EXAMINERS

Chairman

Dr. Sheak Rashed Haider Noori
Associate Professor and Associate Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Abdus Sattar
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Saiful Islam
Senior Lecturer
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

External Examiner

Farid

Dr. Dewan Md. Farid
Professor
Department of Computer Science and Engineering
United International University

DECLARATION

We hereby declare that, this project has been done by us under the supervision of **Md. Sadekur Rahman, Assistant Professor, Department of CSE** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:

Sadik
____

Md. Sadekur Rahman
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised by:

ZHabib
____

Md. Tarek Habib
Assistant Professor
Department of CSE
Daffodil International University

Submitted by:

Thoufiq
____

Md. Thoufiq Zumma
ID: -181-15-10968
Department of CSE
Daffodil International University

Jerin Akther
____

Jerin Akther Munia
ID: -181-15-11136
Department of CSE
Daffodil International University

Dipankar
____

Dipankar Halder
ID: -181-15-11137
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First we express our heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final year project/internship successfully.

We really grateful and wish our profound our indebtedness to **Md. Sadekur Rahman, Assistant Professor**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "*Field name*" to carry out this project. His endless patience ,scholarly guidance ,continual encouragement , constant and energetic supervision, constructive criticism , valuable advice ,reading many inferior draft and correcting them at all stage have made it possible to complete this project.

We would like to express our heartiest gratitude to **Dr. Touhid Bhuiyan, Professor and Head,** Department of CSE, for his kind help to finish our project and also to other faculty member and the staff of CSE department of Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, we must acknowledge with due respect the constant support and patience of our parents.

ABSTRACT

Social networking sites have become the most popular Internet destination, giving social scientists a unique chance to study online behavior. A rising number of study articles on social media are being published, with just a few of them focusing on personality prediction. Personality assessments computer based on data from social media platforms has proven to be more accurate than judgments given by persons who are familiar with the topic. Text on social networking sites is used to automatically detect an individual's personality qualities. We are using the Myers-Briggs Type Indicator (MBTI) dataset. These datasets have 16 types and 8675 posts. From the input text, we categorized four personality qualities using the Myers-Briggs Type Indicator. They are, in particular Introversion-Extroversion (I-E), Intuitions-Sensing(N-S), Feeling-Thinking(F-T), and Judging-Perceiving(J-P) . This data set we collected from Kaggle. Firstly, preprocessing the dataset. This is text data so that we are using NLP for preprocessing. Then using the machine learning techniques. Tokenization, word stemming, stop words deletion and feature, as well as TF IDE, are examples of text preprocessing methods. We are using six machine learning algorithm. We compare all of algorithm, among them the Support Vector Machine (SVM) have best for highest accuracy. Support vector machines outperform the other six machine learning algorithms in terms of accuracy, according to the results of an experiment.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Introduction

Personality is the characteristic sets of behaviors, cognitions, and emotional patterns that evolve from biological and environmental factors. While there is no generally agreed upon definition of personality, most theories focus on motivation and psychological interactions with the environment one is surrounded. So for example, someone who prefers introversion, intuition, thinking and perceiving would be labelled an INTP in the MBTI system, and there are lots of personality based components that would model or describe this person's preferences or behavior based on the label. Most of the work of a personality depends on its predictions. One of these models is mostly used. This is it The Big Five (Goldberg, 1990) is a Well-established model that classifies personality Features with five dimensions: extroversion, consent, Conscience, nervousness and openness. In contrast, the Myers-Briggs type indicator Model (MBTI) (Myers et al., 1990) .There are 16 personality types in four dimensions: Introversion / extroversion (how one gains strength),Sensing / Intuition (the way someone processes information),Thoughts / feelings (how someone decides), and Judging / realizing (how one presents oneself or Yourself in the outside world) despite some controversy About the validity and reliability of the test (Barbuto) Jr., 1997), the MBTI model has been found numerous For applications, especially Art 1 and self-discovery. .Personality tests have attracted increasing attention in recent years, as these became mandatory for most businesses and organizations. Individuals who want to start up a business face a significant challenge. The time and cost of a personality test are 2 factors to take into account. The mass of personality tests involves 50-70 questions, which could be stressful. For the user, it is extremely unpleasant. Our project's focus is to enhance profitability. By limiting the amount of text entered by the user, this process can be finished. It is indeed completely free. Our assessment is based on the Myer test. Briggs Test Indicator (Briggs Test Indicator) (used as MBTI in rest of the sections).This is a very well type of assessment According to them, There are 16 different character types in this assessment. It identifies a personality.

The increasing popularity of social media platforms like Twitter and Facebook has encouraged the online community to share their thoughts, emotions, opinions, and sentiments with others. Other social opinions, behaviors, and characteristics are mirrored in themselves. Certainly, there is indeed a strong link between personal minds. Personality, and also the way they behave on social networking sites in the form of tweets or comments. Researchers are currently interested in developing automatic personality recognition methods, specific to personality identification from social media sites. The philosophy of such applications is based on a variety of elements. Big Five Factor Personality Model is a representation of a model of personality. Our approach, is from the other hand, is based on an available public dataset that contains each user's MBTI personality traits label. The Myers-Briggs Type Indicator (MBTI) is a personality assessment tool based on Carl Jung's type theory. This is still carrying on. It continues to be the most widely used personality test in the world today. MBTI is a type of personality test that is widely available in public databases Indicators of kind.

The Big-five and the MBTI and found that the Agreeableness (A) score from the Big-five is related to the MBTI. Only the Thinking-Feeling model was shown to be connected (T-F).MBTI has a dimension. Conscientiousness (C) was discovered to be a positive trait. Both the Thinking-Feeling (T-F) and Judging-Perceiving (J-P) dimensions are linked. Extraversion (E) was shown to be a prominent personality trait. Extraversion-Introversion (E-I) dimension is associated. The Big-five model's Openness (O) score was all four are linked, notably Sensing-Intuitive (S-N), dimensions. There was no correlation between neuroticism (N) and any MBTI subscale score. Given the association between the MBTI and the Big-five model, our personality's performance will be affected. Other works may be compared to the trait categorization system, which forecasts each user's MBTI personality category. That predict the Big-five personality scores in this domain. The analysis of online human behavior on this massive, ecologically valid dataset is an attempt to understand the complex relationships that underpin social phenomena. The goal of our research is to look at language-based personality trait classification models. We demonstrate with this effort that, that we can learn a lot about the latest 50 tweets just by looking at a small portion of them can reliably identify some personality characteristics

1.2 Motivational

An important relationship and prediction has been identified as a measure of a person's feeling ability or confidence .Through this we can explain the feelings of a person and through this we can explain the behavior of a person. This model can also explain a person's thinking. We can also explain his judgmental analysis if we want. We can use this model to show how someone decides. Through this we can explain the characteristics of a person. Through this model we can get an idea about a person without hearing anything from him/her.

1.3 Relational of the study:

If we want to get an idea of a person, we need to know his characteristics .We need a few days Movement with this person but once we know some of the characteristics of a person through this work, we can give a complete idea of what that person is like through model. Through this work we can easily gain an idea about a person. We'll go over the entire research of our study. We used kaggle to acquire data from the Myers-Briggs Personality Type Dataset (MBTI). Personality tests have gained in popularity in recent years as they have been required in most businesses and organizations.

1.4 Research Questions:

- What is Personality Prediction Analysis?
- How personality Prediction analysis works?
- What are the benefits of Personality Prediction analysis?
- What are the problems of Personality Prediction analysis using machine learning?

1.5 Expected Output

We plan to publish a research article in this relevant sector because this is research and our goal is to go for publication after acquiring a decent accuracy.

Our major objective is to combine all of the different algorithms into one study project in order to determine which machine learning approaches perform better than others. So, we can say that the main distinction between our study and that of other academics is that we are both attempting to discover the best accurate machine learning approach for identifying or predicting personality. We have to use six algorithms in this study project to compare and determine the best algorithm for the best machine learning approach, which will help us in determining or predicting personality. We expect a higher level of precision. For this personality prediction identification, we will utilize machine learning approaches and algorithms that will give us the best accuracy or close to the maximum accuracy. We want to publish this study work in a journal or conference if we improve our accuracy.

1.6 Report Layout

We discussed an overall view of the entire study activity in the first chapter of this paper. The background studies that are pertinent to this subject were covered in the second chapter. The research methodology was explained in the third chapter. We covered which algorithms we utilized in the fourth chapter. The experimental results and conclusion were addressed in the last chapter.

CHAPTER 2

BACKGROUND

2.1 Terminologies

Machine learning is a well-known technology for personality prediction that is widely used by researchers. The researcher can utilize machine learning to study personality traits because of its benefits in learning past data and generating predictions on future data. In psychological science, such an application is also well-known as a personality evaluation instrument. Businesses and recruiters are increasingly investing in machine learning-based personality prediction technology.

Personality tests have grown in popularity in recent years since they have become a requirement for most organizations and businesses. Individuals who wish to take a personality test encounter two major issues: the expense and the length of the test. The majority of personality tests comprise 50-70 questions, which might be lengthy for the user. This test identifies 16 different personality types. It determines a person's four-letter personality type, with each letter denoting a distinct personality preference or propensity. Introversion (I) vs. Extraversion (E), Intuition (N) vs. Sensing (S), Thinking (T) vs. Feeling (F), and Judging (J) vs. Perceiving (P) are the four personality types (P).

Myers-Briggs Type Indicator (MBTI). They characterized a person's personality as a set of characteristics that indicate a person's chances of being distinctive in their behavior, feelings, and thoughts. These characteristics of a person alter with time and under different situations. In a nutshell, personality is a collection of traits and standards that combine to form an individual's distinct personality.

People currently use social media platforms to express their views and emotions. The postings can be made in a variety of ways, such with a picture, a URL link, or music. Social media may also be used to investigate people's personalities. Companies nowadays tend to check candidates' social media accounts in the process of picking the ideal individuals to learn about their personalities for a certain job. They want to cut down on the amount of time spent in the earliest stages of recruiting, which is commonly referred to as social media mining. We used Twitter in this article since it is one of the most prominent social media sites today. Employers find it difficult to choose the finest applicants for their businesses. Furthermore, the conventional approach normally requires companies to spend time interviewing all of the individuals who have been shortlisted.

2.2 Related Works

An author proposes a technique for analyzing a person's social media posts/tweets and generating a personality profile as a result. The focus of the research is on data collecting, pre-processing methods, and prediction using a machine learning algorithm [1]. Different feature selection approaches, such as Emolex, LIWC, and TF/IDF, are used to create the feature vectors. The feature vectors obtained are utilized in the training and testing of several machine learning algorithms, such as Neural Net, Nave Bayes, and SVM. The MBTI dataset, It is based on the social media network Reddit, was presented in [2] for personality prediction. A large number of characteristics are retrieved, and benchmark models for personality prediction are examined. SVM, Logistic Regression, and other techniques are used to classify the data (MLP). Across all MBTI aspects, the classifier that included all language characteristics excelled. From tweets made in English and Indonesian, a method was constructed to detect user personality using the Big Five Factor personality model [3]. On the My Personality dataset, many classifiers are used. Naive Bayes (NB) has a 60 percent accuracy, which is higher than KNN (58 percent) and SVM (50 percent) (59 percent). Although the accuracy of prior study (61%) was not improved, the aim of determining personality from Twitter-based posts was attained. The results might be improved by using a larger dataset and applying a semantic approach.

Existing work on social media text personality detection relies on supervised machine learning approaches applied to benchmark datasets [4], [5]. The skewness of the datasets, i.e. the presence of unbalanced classes with regard to distinct personality characteristics, is, however, the fundamental concern with the aforementioned research. This problem is mostly to blame for the decline in performance of the personality recognition system. [6] Developed an unsupervised personality categorization approach to highlight the issue of how various personalities interact and behave on the social networking platform Twitter.

This study employs linguistic and statistical features, which are subsequently evaluated on a data corpus clarified using a personality model based on human evaluation. In [7], the author suggested a personality detection system based on the Big-5 personality model that uses an unsupervised method. For the extraction and categorization of user attributes, many social media network sites are employed. Linguistic characteristics are used to create a personality model. For an input text, the algorithm predicted personality and produced respectable results. Extended annotated corpus, on the other hand, can improve the system's performance. Despite the fact that supervised machine learning algorithms have been used to measure personality [8, 9]. For MBTI personality evaluation, the state-of-the-art Algorithm XGBoost with optimal parameters is utilized [10]. When compared to other machine learning algorithms, the XGBoost classifier produces higher accuracy [10, 11]. The proposed study is the first to use XGBoost as a classifier and the MBTI as a personality model to predict personality from text. The Big Five personality characteristics have been linked to user behavior on social media. Certain behavioral patterns found on social media sites can be used to describe each of the five qualities [12]. Individuals' levels of neuroticism and agreeableness have been linked to the frequency with which they use negative language in their postings and their proclivity to use swear words or express pleasant emotions in their posts [13].

2.3 Research Summary

Table 2.1: Reference paper analysis and summary

Author	Approach	Performance
Pratama et al.	Using supervised Algorithm KNN,NB,SVM	Accuracy KNN =58%,NB=60%,SVM=59%
Bharadwaj et al.	TF-IDF, Emolex, LIWC, SVM, Neural Net, and Nave Bayes as well as ConceptNet	SVM with all feature vectors had the greatest accuracy in all MBTI dimensions.

Plank ei al.	Logistic regression is a technique for predicting the outcome of The binary word n-gram is employed as a feature selection in the model.	percent accuracy in personality prediction I/E= 72.5% S/N = 77.5% T/F =61.2% J/P = 55.4%
Golbeck et al.	ZeroR and GP are the two regression models applied.	Open has a higher accuracy rate of 75.5 percent, whereas Neuro has a lower accuracy rate of 42.8 percent.

Our research work we have the goal to use a dataset to predict personality prediction. We get the dataset from kaggle. Firstly we are preprocessing our dataset. Using NLP and then implement different types of machine learning algorithm. According to different model and techniques are for each of them. And we will try 6 different machine learning algorithm. This algorithm are detect personality prediction. All are algorithm result are compare and we take best accuracy which is give us the best prediction.

2.4 Scope of the Problem

There is a lot of work about personality prediction and there is much research is used in a data set that is collected by Twitter. This dataset is used in many ways they are using many algorithms and some time that they cannot get the best accuracy .this dataset was text data so that we preprocess it by natural language process than using machine learning model finally we find out this dataset getting a perfect result for best accuracy that's why we are using this dataset to find the best accuracy of the Other researcher.

2.5 Challenges

There are some challenges we have to face while doing this research work. Our dataset has 2 columns, there are type and posts. All datasets are string types. And posts are text files. We know that machines cannot detect text files. So that we're preprocessing our text file by Natural Language Preprocessing. We are using stop words, tokenization, stemming. We are removing extra spaces, removing punctuation. Removing preceding spaces before common punctuation such as a full stop, comma, or question mark. For preprocessing removing URLs link, also remove non-words and taking only words. Removing multiple letters repeating words and transforming MBTI to binary vector and binary MBTI personality into 4 letters. Run,

Then we have to process the data and change some value according to our research work because we have to prepare the data set for multiple machines learning algorithms. As we already mentioned that we are trying to use six different machine learning algorithms in one research work that means we have to prepare the data set for all those six machines learning algorithms.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Research Subject and instrumentation

In this first place, we have discussed the process of detecting Personality prediction with the conceptual and theoretical process. Machine learning model and NLP models needs high configuration pc with GPU and others instruments. Now a list given below of the required instruments for this model.

Hardware and Software:

- Intel Core i5 8th generation
- 1 TB Hard Disk Drive
- 8GB RAM

Development Tools:

- Windows 10
- Python 3.8
- Pandas
- Seaborn
- Matplotlib
- Numpy
- Scikit-Learn
- NLTK

3.2 Data collection Procedure

We have collected Dataset from kaggle. Various punctuation marks and numerous emoticons were also included in the data, which had to be cleaned and eliminated. The data were subjected to several rounds of scrutiny. All superfluous punctuations are removed from the data. Emoticons and "stop words" such as "a," "the," and "the" were eliminated.

This was accomplished with the use of regular expressions in Python and the Text processing library NLTK. This strategy was comparable to that of is an example of this. With simply a few lines of code, a clean data collection was created. Statements that were useful in understanding the situation Personality characteristics some personality types, according to the MBTI, are more widespread than others. As a result, there is an excess of a subset of the dataset's personality types.

3.3 Implementation and Algorithms

3.3.1 Random Forest Classifier

Random forests is a method for supervised learning. It has the ability to be utilized for both classification and regression. It's also the most adaptable and user-friendly algorithm. The trees make up a forest. A forest is thought to be stronger the more trees it has. Random forests generate decision trees from randomly chosen data samples, obtain predictions from each tree, and then vote on the best option. It also serves as a strong indicator of the value of the feature.

Random forests may be used for a range of tasks, including recommendation engines, image classification, and feature selection. It may be used to identify fraudulent activities, categorize loyal loan applicants, and anticipate illnesses. The Boruta method, which picks relevant characteristics in a dataset, is built on it.

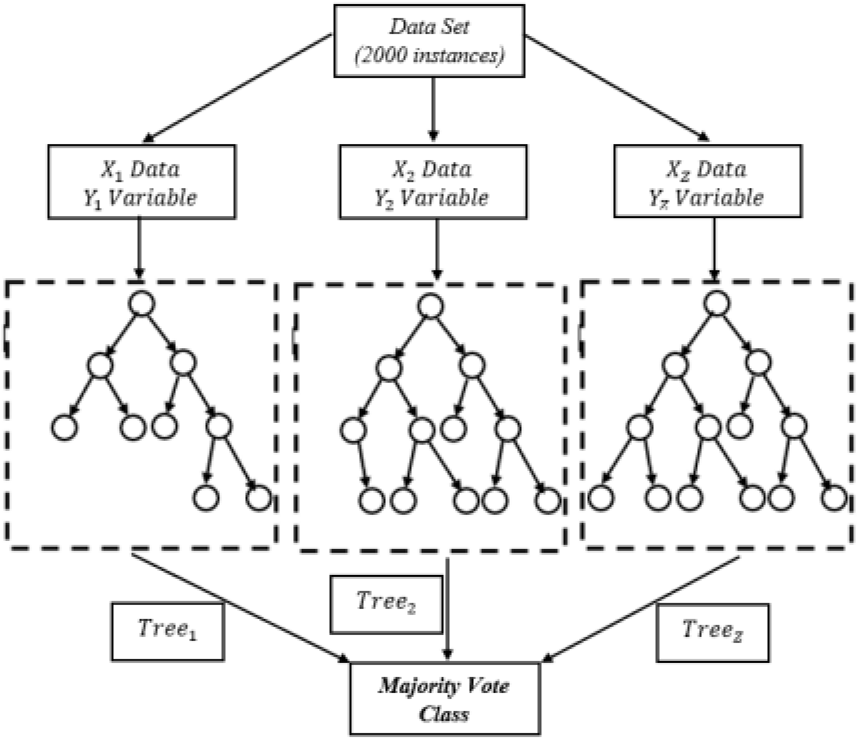


Figure 3.1: Steps of how random forest classifier works

There are four steps to it:

- Choose random samples from a set of data.
- Create a decision tree for each sample and use it to generate a prediction result.
- Make a vote for each expected outcome.
- As the final forecast, choose the prediction with the most votes.

Because of the large number of decision trees involved in the process, random forests is regarded a very accurate and resilient approach. It is not affected by the problem of overfitting. The fundamental reason for this is that it averages all of the forecasts, canceling out any biases. Both classification and regression problems can benefit from the approach. Missing values can also be handled using random forests. There are two methods for dealing with missing data: utilizing median values to replace continuous variables and finding the proximity-weighted average.

The relevance of each feature is calculated using gini importance or mean decrease in impurity (MDI) in random forest. The entire decrease in node impurity is also known as Gini significance. When you remove a variable, this is how much the model's fit or accuracy suffers. The more significant the variable, the greater the drop. The mean decrease is an important metric for variable selection in this case. The Gini index can be used to describe the variables' total explanatory power.

3.3.2 XGBoost Model

These days, XGBoost is one of the most widely used machine learning algorithms. Regardless of whether the goal at hand is regression or classification,other machine learning methods are known to offer greater results than XGBoost. Indeed, it has become the "state-of-the-art" machine learning technique for dealing with structured data since its beginnings.

The gradient boosting (GBM) architecture is at the heart of XGBoost (Extreme Gradient Boosting), a family of boosting algorithms. It's a distributed gradient boosting library that's been optimized.

Boosting is a sequential strategy that is based on the ensemble concept. It combines a group of ineffective learners to increase prediction accuracy. The model outcomes are weighted at any instant t depending on the results of the preceding instant t-1. The outcomes that were accurately predicted are given a lower weight, whereas those that were misclassified are given a larger weight. A weak learner is one who is only marginally better than guessing at random.

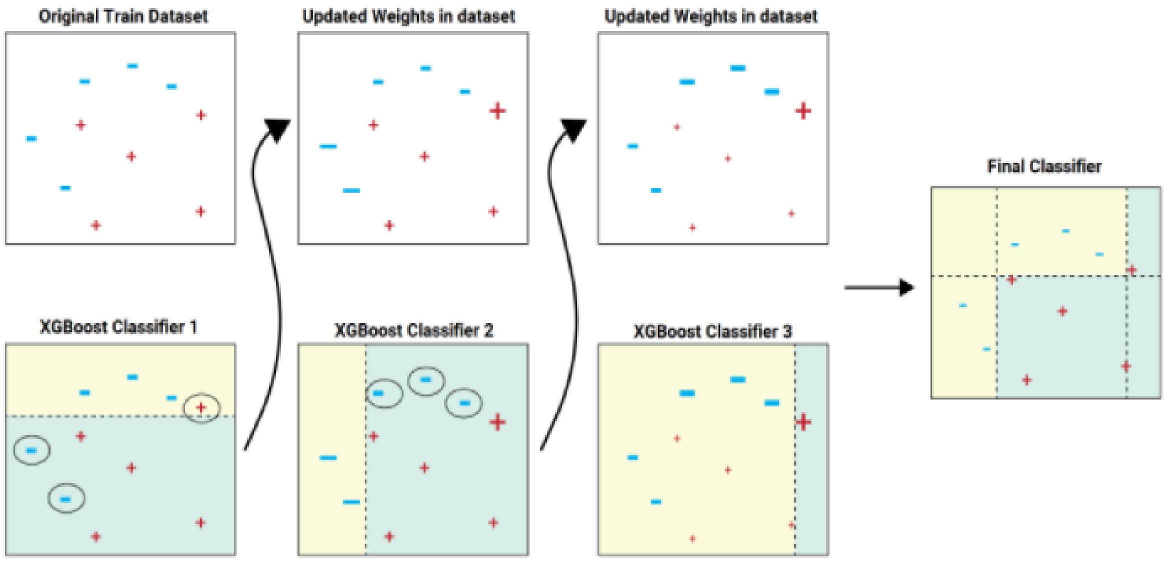


Figure 3.2: Steps of how XGBoost classifier works

The train dataset is supplied to the classifier 1 in the example above. The classifier predicted hyphen with a yellow backdrop, whereas it predicted plus with a blue background. The classifier 1 model predicts two hyphens and one plus wrongly. A circle has been placed over them. The weights of these data points that were mistakenly predicted are raised, and they are transmitted to the next classifier. To classifier 2, that is. Classifier 2 successfully predicts the two hyphens that classifier 1 couldn't. However, classifier 2 makes a few more mistakes. This approach is repeated until we have a combined final classifier that correctly predicts all data points.

The classifier models can be added until all of the objects in the training dataset have been properly predicted, or until the maximum number of classifier models has been reached. Hyper parameter tweaking may be used to identify the appropriate maximum number of classifier models to train.

3.3.3 SGD Classifier

Let's define Gradient Descent first before moving on to Stochastic Gradient Descent (SGD). Gradient Descent is a well-known optimization strategy in Machine Learning and Deep Learning, and it may be used for nearly all learning algorithms. A function's gradient is its slope. It determines how much a variable changes in reaction to changes in another variable. Gradient Descent is a mathematically defined convex function whose output is the partial derivative of a collection of input parameters. The higher the slope, the greater the gradient.

If you have a million samples in your dataset, you will have to utilize all of them to complete one iteration of the Gradient Descent, and you will have to do this for each iteration until the minima are achieved if you use a traditional Gradient Descent optimization approach. As a result, it becomes computationally prohibitively costly to carry out.

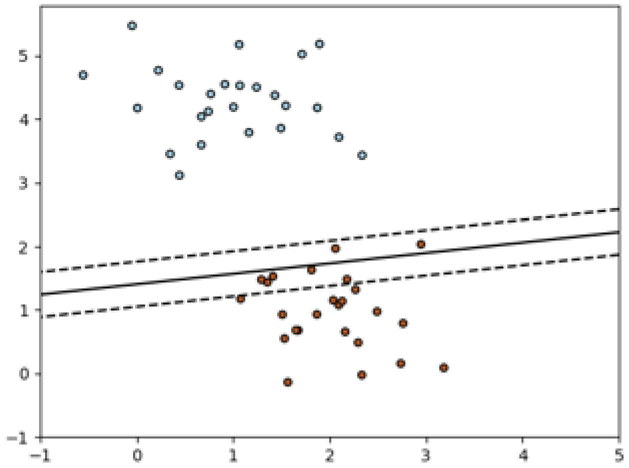


Figure 3.3: SGD Classifier works

Stochastic Gradient Descent is used to tackle this problem. Each iteration of SGD is performed with a single sample, a batch size of one. The sample is jumbled and chosen at random to execute the iteration.

3.3.4 Logistic Regression

A supervised classification technique, logistic regression, is. In a classification problem, the goal variable (or output), y, can only take discrete values for a given set of features (or inputs). Contrary to popular belief, logistic regression is a regression model. The model generates a regression model to predict the chance that a given data input would fall into the category "1." Similar to linear regression, logistic regression models data using the sigmoid function, assuming that the data follows a linear distribution. As a general linear model, we can say that. Equation of linear regression.

$$y = \beta 0 + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots\ldots + \beta_n X_n$$

Sigmoid function : $\quad p = \frac{1}{1+e^{-y}}$

Final equation after applying sigmoid function :

$$P = \frac{1}{1+ e^{-(0+\beta_1 X_1+\beta_2 X_2+\cdots\ldots\ldots+\beta_n X_n)}}$$

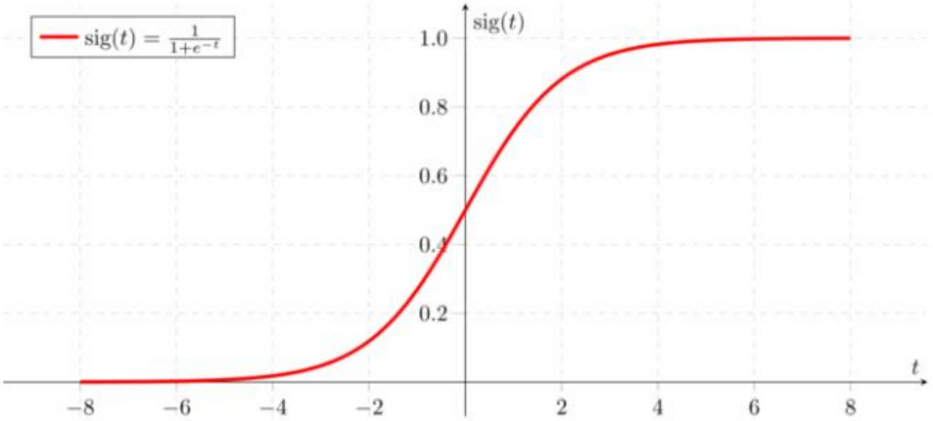


Figure 3.4: Logistic Regression

When a decision threshold is included, logistic regression transforms into a classification procedure. Setting the threshold value is a crucial part of Logistic regression, and it is determined by the classification issue.

The accuracy and recall levels have a significant influence on the threshold value determination. In an ideal world, both accuracy and recall would equal 1, but this is rarely the case.To determine the threshold, we consider the following factors: -

1. Low Precision/High Recall: We pick a decision value with a low Precision or a high Recall in situations where we wish to lower the number of false negatives without necessarily reducing the number of false positives.
2. High Precision/Low Recall: We pick a decision value with a high Precision or a low Recall in situations where we wish to lower the number of false positives without necessarily reducing the number of false negatives.

3.3.5 KNN Model

The K-Nearest Neighbour method is based on the Supervised Learning approach and is one of the most basic Machine Learning algorithms. It is easy-to-implement technique that may be used to address both classification and regression issues. The K-NN approach assumes that the new case/data and old cases are comparable and places the new case in the most similar category to the existing categories. The K-NN approach maintains all available data and classifies new data points based on their similarity to previous data. This means that utilizing the K-NN approach, new data may be swiftly sorted into a well-defined category.

Although the K-NN method may be used for both regression and classification, it is more typically employed for classification. The K-NN algorithm is a non-parametric algorithm, meaning it doesn't make any assumptions about the data. It's also known as a lazy learner algorithm since it doesn't learn from the training set straight away; instead, it keeps the dataset and uses it to categorize it later. The KNN method merely saves the dataset throughout the training phase, and when it receives new data, it classifies it into a category that is quite similar to the new data.

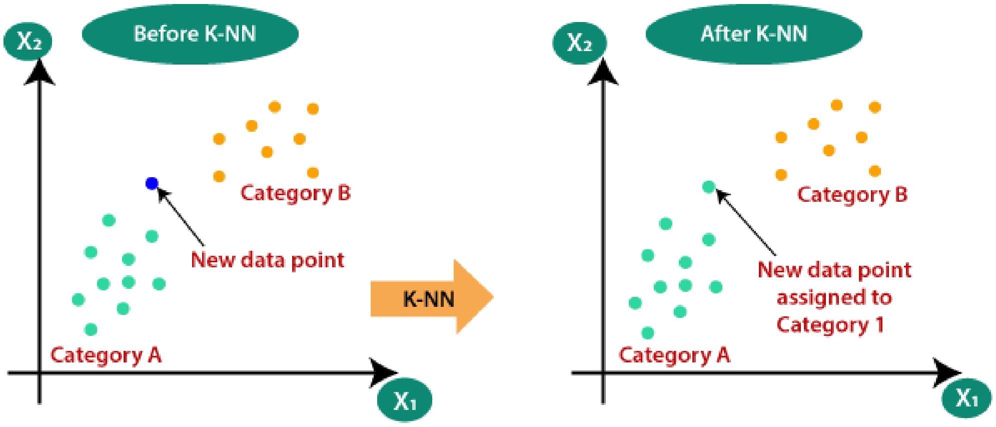


Figure 3.5: KNN model works

We run the KNN algorithm numerous times with different values of K to find the K that decreases the amount of mistakes we encounter while retaining the algorithm's capacity to generate correct predictions when it's given data it hasn't seen before.

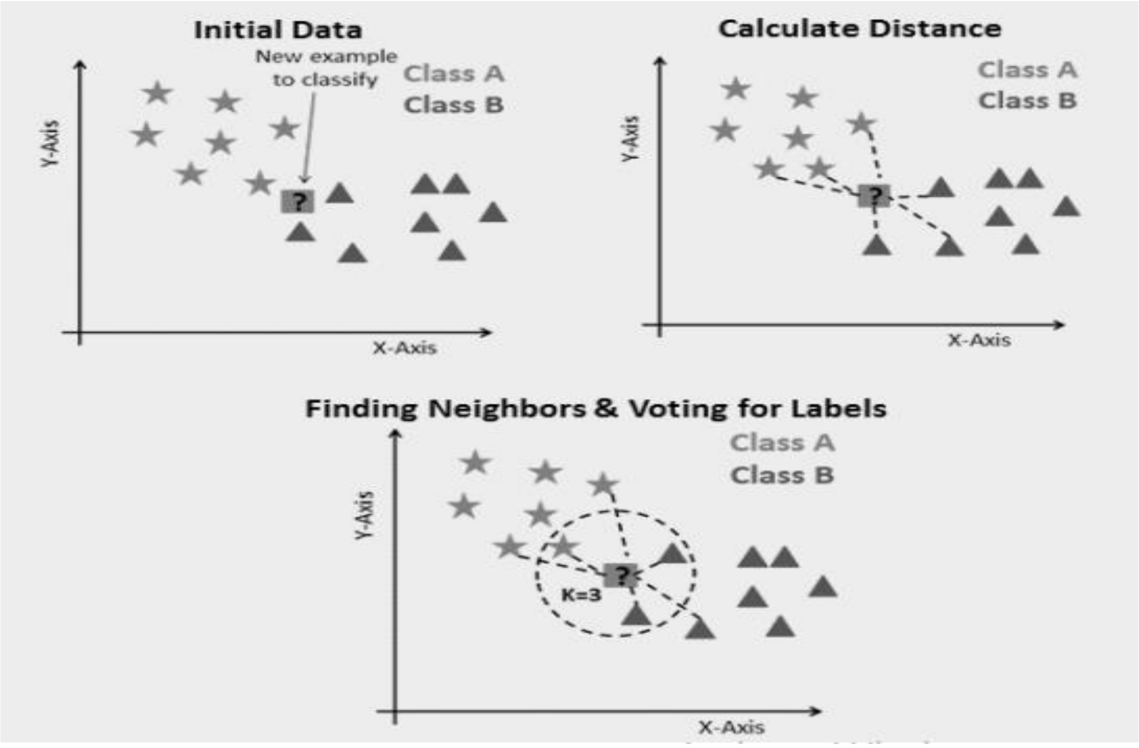


Figure 3.6: The process of KNN algorithm

K as a regulating variable in a projection model. For all data sets, there are no ideal neighbors. Each dataset has its own set of requirements. The vibration will have a higher influence on the findings for a small number of neighbors, making it computationally expensive for a large number of neighbors. According to research, the smoother judgment limit with a limited number of nodes is the most flexible match, with low bias but high variation, whereas a big number of neighbors suggests lower variance but higher bias. When the number of parties equals two, data scientists usually obtain an odd number.

$$\mathrm{d(p,q)} = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \ldots\ldots (q_n - p_n)^2}$$

$$\sqrt{\sum_{i=1}^{n} (q_l - p_i)^2}$$

3.3.6 SVM Model

The Support Vector Machine, or SVM, is a common Supervised Learning technique that may be used to solve both classification and regression issues. However, it is mostly utilized in Machine Learning for Classification difficulties.

The SVM algorithm's purpose is to find the optimum line or decision boundary for categorizing n-dimensional space into classes so that additional data points may be readily placed in the proper category in the future. A hyperplane is the name for the optimal choice boundary. The extreme points/vectors that assist create the hyperplane are chosen via SVM. Support vectors are the extreme instances, and the method is called a Support Vector Machine. Consider the image below, which shows the usage of a decision boundary or hyperplane to categorize two distinct categories:

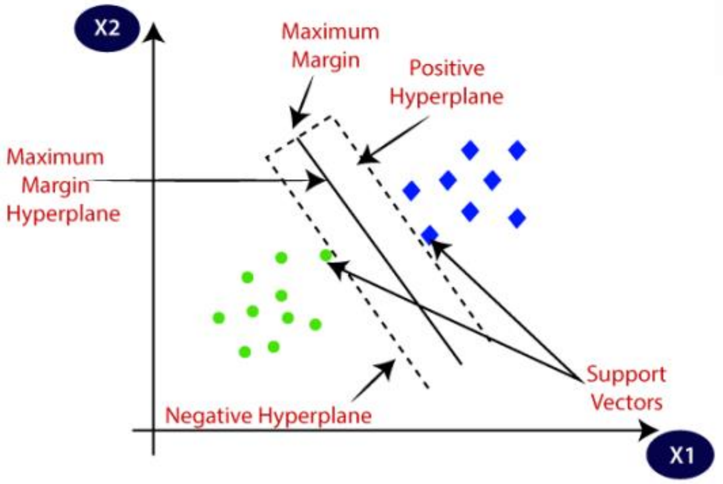


Figure 3.7: SVM figure

There are two types of SVM:

Linear SVM: Linear SVM is a classifier that is used for linearly separable data, which implies that if a dataset can be categorized into two classes using a single straight line, it is called linearly separable data, and the classifier is named Linear SVM.

Non-linear SVM: Non-linear SVM is used for non-linearly separated data, which implies that if a dataset can't be categorized using a straight line, it's non-linear data, and the classifier employed is called Non-linear SVM.

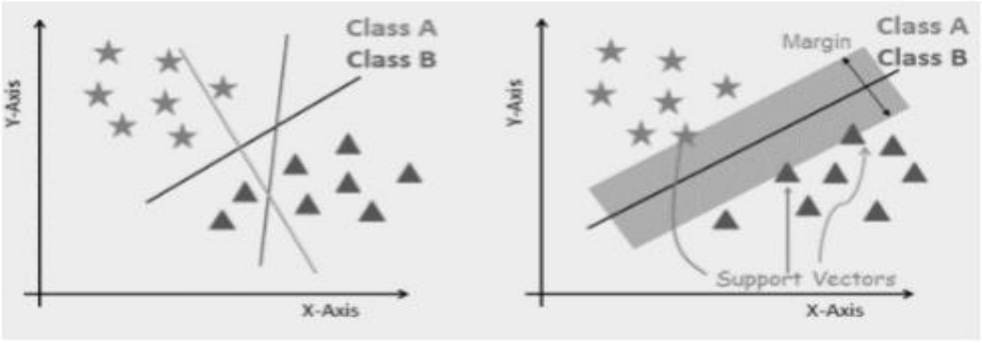


Figure 3.8: How SVM works

There are several issues with linear hyperplanes that cannot be addressed when dealing with nonlinear and indistinguishable planes, as seen below (left-hand side).In that case, SVM employs a kernel technique to convert the establishment space into a higher level. As illustrated on the right, three-dimensional feature space. The data points are plotted on the x- and z-axes.(Z is the squared number of x and y: z=x2=y2) are presented. With linear segregation, this is no longer the case. These points are easily distinguishable.

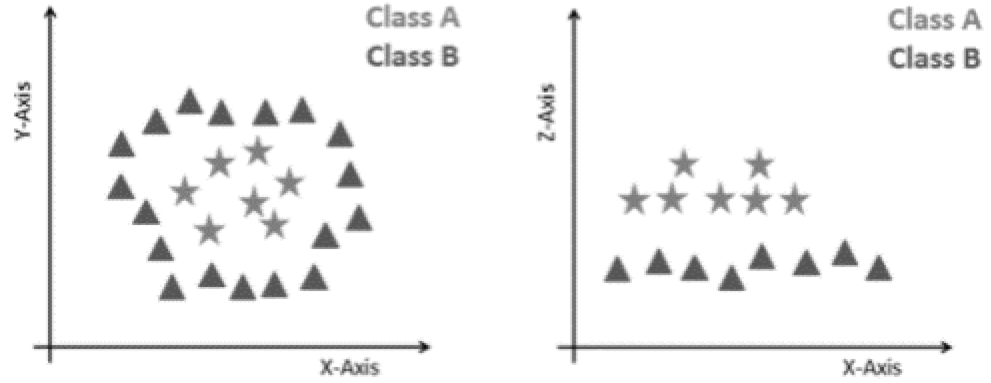


Figure 3.9: Managing nonlinear and indistinguishable planes

The SVM algorithm is implemented using the kernel. A kernel can convert the data input space into the required format. SVM employs a kernel trick technique. The kernel in this scenario takes a limited input area and makes it larger. To put it another way, it's possible. It is claimed that adding more dimensions changes a seemingly unrelated situation into difficulties that can be separated it’s very beneficial in non-linear separation difficulties. The technique with the kernel makes it possible to create a more thorough categorization.

Any two actions can be utilized as a linear kernel in a conventional dot product. The total of each pair of input data multiplied equals the product of the two vectors. This is how the mathematical representation looks:

$$\mathbf{K(x,}\boldsymbol{x_i}\mathbf{) = sum (x \times }\boldsymbol{x_i}\mathbf{)}$$

A more common linear kernel structure would be the polynomial kernel. The nucleus Polynomial will distinguish between curved and nonlinear input spaces.

$$\mathbf{K\ (\ x\ ,\ \boldsymbol{x_i}) = 1 + sum(x\ \times\ \boldsymbol{x_i})^{\boldsymbol{d}}}$$

The polynomial stage is denoted by d. d=1 is a near approximation in practice. The degree must be explicitly specified in the learning method. The radial functional kernel is a well-known kernel function that aids in vector machine classification. RBF may remap input space in infinite dimensional space.

$$\mathbf{K\ (x,\boldsymbol{x_i}) = exp(-\mu \times sum(x\ -\ \boldsymbol{x_i})^2}$$

Gamma is a function that ranges from 0 to 1. A larger gamma value corresponds to the training dataset, causing unneeded adaptation. Gamma=0.1 is often considered to be an excellent expected value. The gamma value must be defined throughout the learning method. Separately. SVM categories provide reasonable results when compared to the Naive Bayes method. Precision, and make more accurate predictions. They frequently employ less memory and make decisions based on a subset of instructional points. SVM works best when there is a lot of separation.

3.4 Proposed Methodology

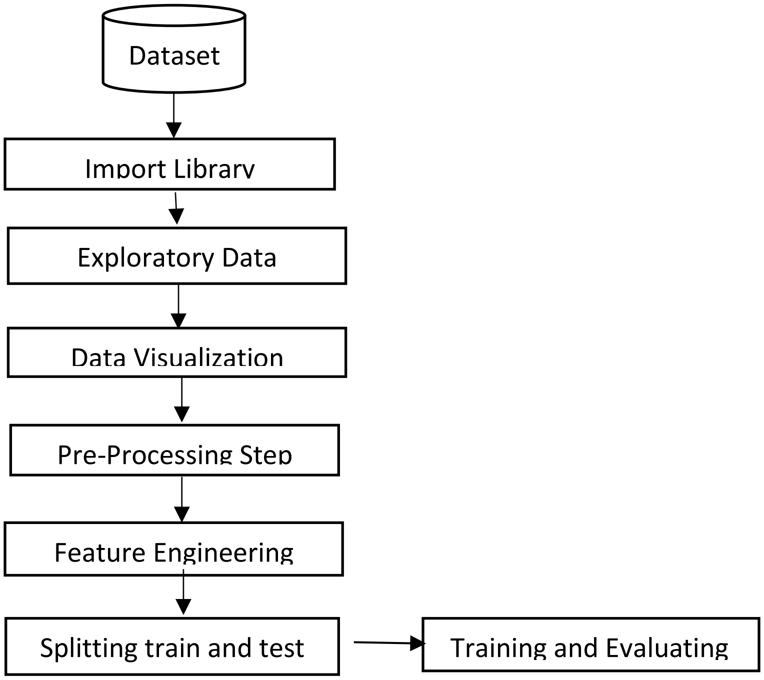


Figure 3.10: Workflow for MBIT

3.4.1 Get Dataset

As previously we have that we have collected data from Kaggle. There are a lot of dataset in the kaggle. We want to work with human personality prediction that’s why we are selected kaggle platform and using MBTI data set. MBTI dataset are used by many way .We want to use several algorithm to detect or predict personality prediction. Here we are added some extra algorithm and improving prepressing step.

3.4.2 Import Library

In this project, here using many library .Like as

- Data analysis library,
- Data visualization library,
- Text Processing,
- Machine learning packages,
- Model training and evolution,
- Some machine learning which are using algorithm.

3.4.3 Exploratory Data Analysis

Firstly loading dataset and using many command lines for exploratory data analysis. Two columns type and post this dataset have 16 type and 8675 posts. There are no null values in our dataset we are checking for null values. Our dataset are object type and there are no duplicate value.Here is type and posts counts.

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

3.4. 4 Data Visualization

We are using here some visualization**.**

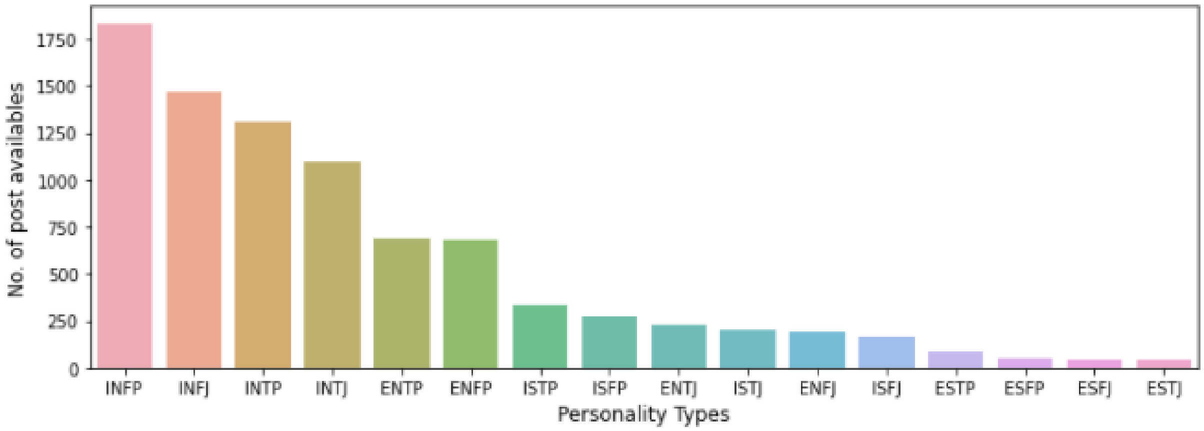
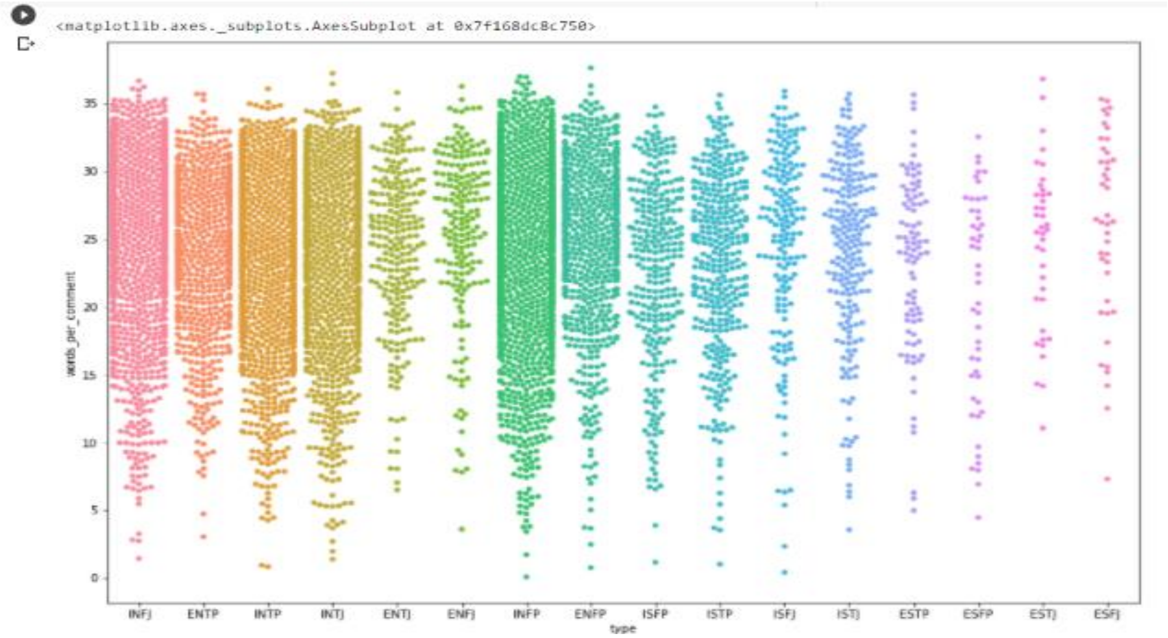


Figure 3.11: Distribution of personality types

In this visualization, plotting this in descending order for better understand. The original dataset only came 2 feature, per type 50 posts. Above visualization we understand that how many posts have each type.



```
<matplotlib.axes._subplots.AxesSubplot at 0x7f168dc8c750>
```

Figure 3.12: The Number of imbalances

The number of imbalances in our dataset is depicted in this graph. We can plainly see that INFP has the most congested plot, indicating that there are the most comments of this personality type. We know That Imbalanced dataset is relevant primarily in the context of supervised machine learning involving two or more classes. Imbalance means that the number of data points available for different classes is different: If there are two classes, then balanced data would mean 50% points for each of the classes.

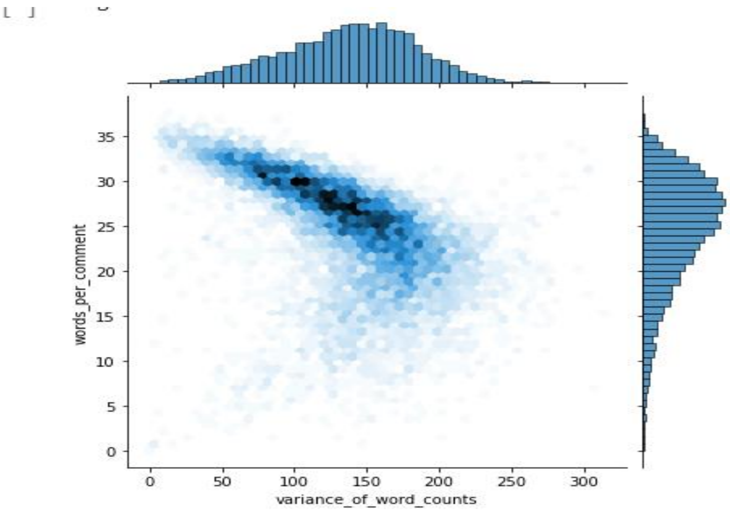


Figure 3.13: Joint plot of word comments

A Joint plot is made up of three plots. One plot depicts a bivariate graph that demonstrates how the dependent variable (Y) fluctuates with the independent variable (X) (X). Another plot, located horizontally at the top of the bivariate graph, depicts the independent variable's distributions (X) .We utilized a plot called connect plot. It is a seaborn library that may be used to visualize and analyze data fast. The variation of word counts and the number of words per remark are shown in this graph.

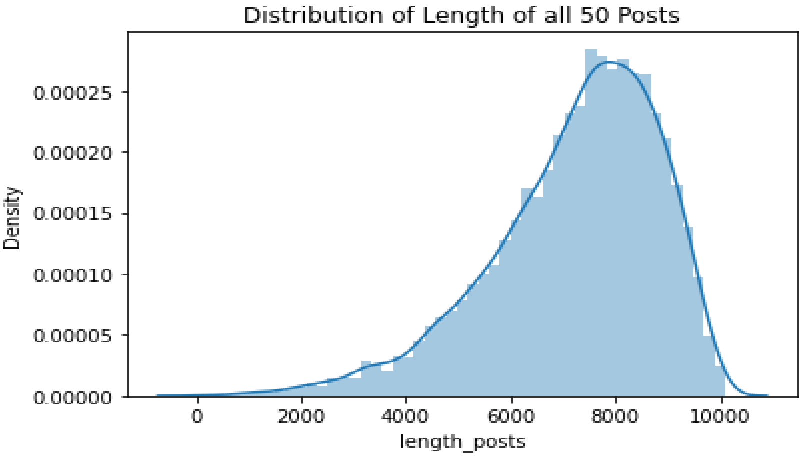


Figure 3.14: Distribution of length

A histogram is a graph that allows you to identify and display the underlying frequency distribution (shape) of continuous data. This enables data to be examined for its underlying distribution (e.g., normal distribution), outliers, skewness, and other factors. We're utilizing histogram graphs in this example. The majority of long posts are between 7000 and 9000 words long.

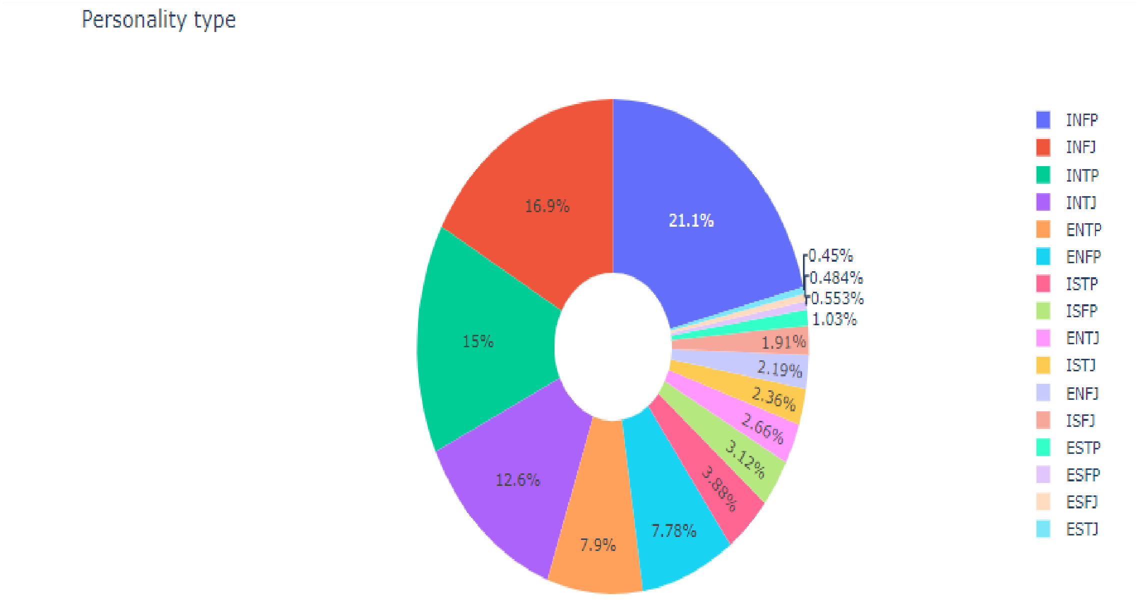


Figure 3.15: Pie Chart

A pie chart is a form of a graph that uses a circular graph to represent data. The graph's parts are proportionate to the percentage of the total in each group. In other words, the size of each slice of the pie is proportional to the size of the group as a whole. Here's a pie chart that shows how many different categories are contained in this dataset.

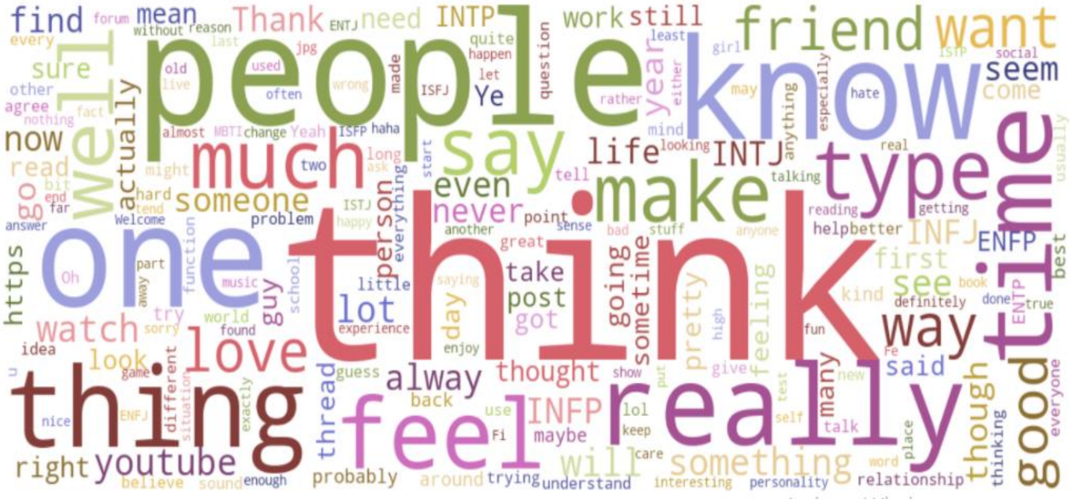


Figure 3.16: Word cloud

The term "word cloud" refers to a graphic depiction of the frequency of words. The larger the keyword appears in the graphic created, the more frequently it appears in the text being analyzed. Word clouds are becoming more popular as a simple technique for determining the topic of written text. A Word Cloud, also known as a Tag Cloud, is a visual representation of text data in the form of tags, which are often single words whose significance is represented by their size and color. You may use the same method to evaluate data to any other source, including Twitter and Facebook.

There are a lot of terms in this MBIT dataset. A word cloud is a type of data visualization that displays our most frequently used words. In addition, we are creating 16 word cloud visualizations for 16 personality types in our coding notebook.

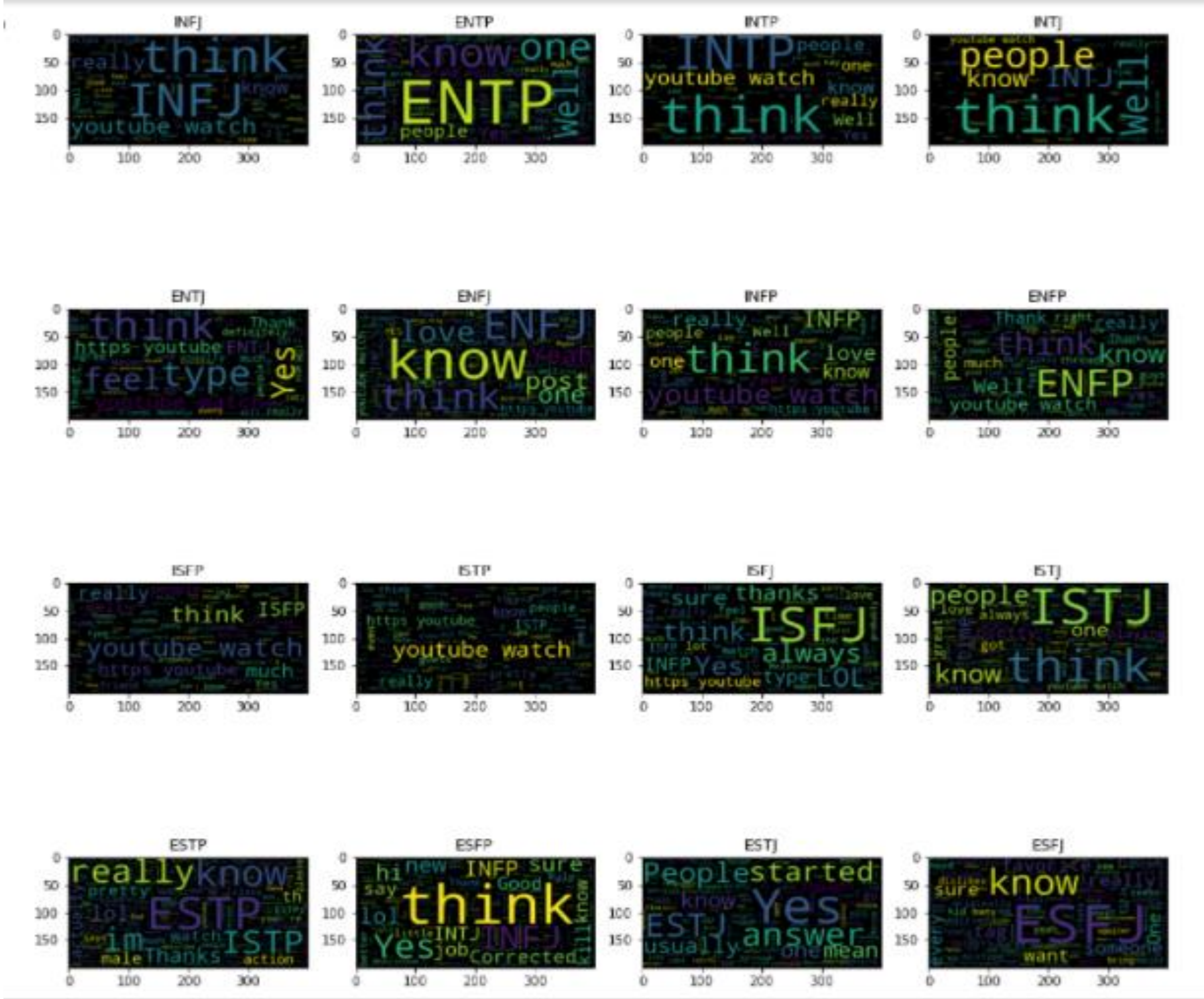


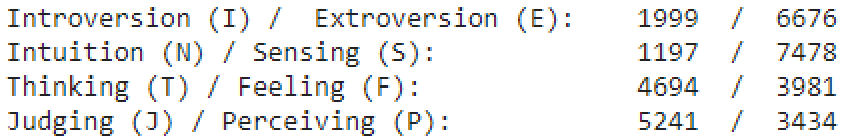
Figure 3.17 personality Word cloud

3.4.5 Data Pre-Processing

We grouped each of the 16 types into four groups.

- Introversion (I) – Extroversion (E)
- Intuition (N) – Sensing (S)
- Thinking (T) – Feeling (F)
- Judging (J) – Perceiving (P)

There are four axes: IE, NS, TF, and JP. It'll either be 1 or 0.



```
Introversion (I) /  Extroversion (E):    1999  /  6676
Intuition (N) / Sensing (S):             1197  /  7478
Thinking (T) / Feeling (F):              4694  /  3981
Judging (J) / Perceiving (P):            5241  /  3434
```

Figure 3.18: personality Word cloud

Here we see that the all post are divided by axis.IE: The highest post is E, and NS: The lowest post is N. We're going through a lot of processes in this preprocessing step. Because we know that machines cannot interpret text or comments, we need first preprocess our data. After that, we deal with it. Natural Language Processing is used in this case.

- Import NLTK it helps out to preprocessing.

The Natural Language Tool kit (NLTK) is a Python-based set of modules and applications for symbolic and analytical natural language processing (NLP) for English. Classification, tokenization, stemming, tagging, parsing, and semantic reasoning are all supported by NLTK.

One of the most popular systems for working with linguistic data is NLTK. Provides APIs for a wide range of text preparation algorithms that are simple to use. It features a huge and engaged community that supports and enhances the library. For Windows, Mac OS X, and Linux, it's free and open-source. Text preparation has long been a crucial step in natural language processing (NLP). It converts text into an easier-to-understand format, allowing machine learning algorithms to perform better.

- Lemmatization

Working with words according to their root lexical components is called lemmatization in natural language processing. It's utilized in computer programming and artificial intelligence for natural language processing and interpretation. The computer may group together words that do not have the same stem but have the same inflected meaning in lemmatization, which is a bit more complicated. Lemmatization is the process of grouping words like "good" with terms like "better" and "best."

The computer may group together words that do not have the same stem but have the same inflected meaning in lemmatization, which is a bit more complicated. Lemmatization is the process of grouping words like "good" with terms like "better" and "best." Lemmatization is also a sign of increasing artificial intelligence complexity, as natural language processing improves its ability to read inputs and produce intelligent outputs as it accommodates lemmatization. As technology becomes closer to passing the Turing test with machines that hear, comprehend, think, and talk like humans, this will become an increasingly significant part of NLP. Run is the lemma of all these words, for example, because runs, running, and ran are all versions of the word run.

- Stop Words

Stop words are words that contribute little sense to a statement in any language. They may be safely ignored without jeopardizing the sentence's meaning. These are some of the most frequent, short function terms for various search engines, such as the, is, at, which, and on.

A stop word is a widely used word (such as "the," "a," "an," or "in") that a search engine has been configured to disregard while indexing and retrieving results as the result of a search query. Stop words include phrases like "if," "but," "we," "he," "she," and "them." We can generally eliminate these terms without affecting the meaning of a text, and doing so enhances a performance of the model (but not always).We don't want these terms to eat up important processing time or take up space in our database. We can simply eliminate them by keeping a list of terms that you regard to be stop words.

- Cleaning some data of the post

The majority of the data scraped from the website is in the form of raw text. Before evaluating or fitting a model to this data, it must be cleaned. To emphasize the qualities that you'll want your machine learning system to pick up on, you'll need to clean up the text data.

- Removing Extra Spaces

Extra spaces in between the words, following or before a phrase, are common in text data. So, to begin, we'll use regular expressions to delete the excess spaces from each phrase.

- Removing Punctuations

The punctuation in the text adds no value to the information. When punctuation is added to any word, it becomes difficult to distinguish it from other words.

- Case Normalization

We just change the case of all characters in the text to upper or lower case in this scenario. Python is a case-sensitive language, therefore NLP and nlp will be treated differently. Using str.lower () or str.upper (), one can quickly transform the string to lower or upper case ().

- Remove stop words

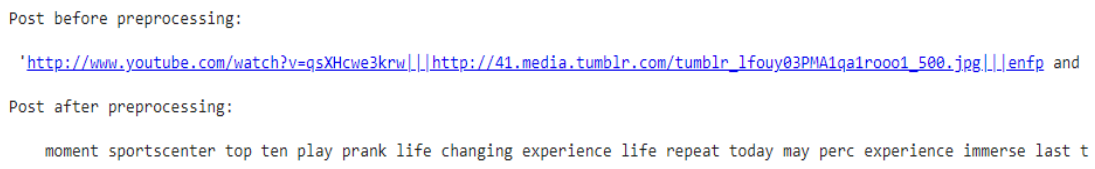
Stop words include: I, he, she, and, but, was, were, being, having, and so on, all of which add no value to the data. As a result, these terms must be deleted, reducing the number of characteristics in our data. After tokenizing the text, they are eliminated.

- Remove white spaces

All gaps should be trimmed such that each word has only one space between them. The conversion should be done in-place, with the solution handling following and leading spaces as well as removing preceding spaces before common punctuation such as a full stop, comma, or question mark.

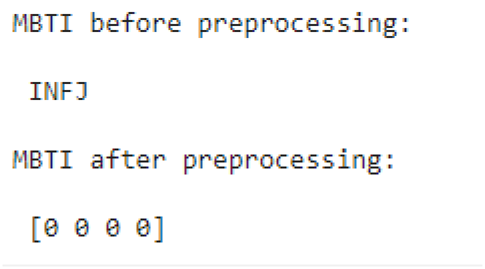
- Remove others things

We should remove urls link, also remove non words and taking only words. Remove Multiple Letter repeating words, Transform MBIT to Binary Vector. Splitting the MBIT personality into 4 letter and binarizing it.



```
Post before preprocessing:

 'http://www.youtube.com/watch?v=qsXHcwe3krw|||http://41.media.tumblr.com/tumblr_lfouy03PMA1qa1rooo1_500.jpg|||enfp and

Post after preprocessing:

    moment sportscenter top ten play prank life changing experience life repeat today may perc experience immerse last t
```

Figure 3.19: After and before preprocessing

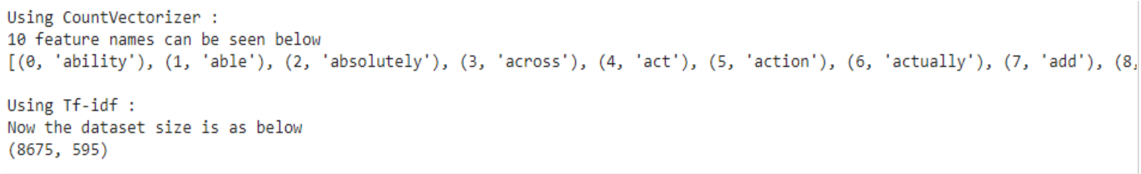


```
MBTI before preprocessing:

 INFJ

MBTI after preprocessing:

 [0 0 0 0]
```

Figure 3.20: Transform mbti to binary vector

3.4.6 Feature Engineering

Feature extraction is essentially a dimensionality reduction method in which raw data is sorted into manageable groupings. The fact that these enormous datasets include a great number of variables, as well as the fact that these variables demand a lot of computational resources to analyze, is a distinguishing trait. As a result, Feature Extraction might be effective in this scenario in terms of picking certain variables as well as merging some of the related variables to decrease the amount of data. Precision and recall metrics would be used to evaluate the data obtained. PCA is a linear dimensionality reduction approach that is widely utilized. It's an algorithm for unsupervised learning.



```
Using CountVectorizer :
10 feature names can be seen below
[(0, 'ability'), (1, 'able'), (2, 'absolutely'), (3, 'across'), (4, 'act'), (5, 'action'), (6, 'actually'), (7, 'add'), (8,

Using Tf-idf :
Now the dataset size is as below
(8675, 595)
```

Figure 3.21: Using countvectorizer and Tf-idf

The scikit-learn toolkit in Python has a fantastic utility called CountVectorizer. It is used to convert a text into a vector based on the frequency (count) of each word that appears throughout the text. This is useful when dealing with a large number of such texts and converting each word into a vector (for using in further text analysis).

CountVectorizer builds a matrix with each unique word represented by a column and each text sample from the document represented by a row. The count of the term in that particular text sample is the value of each cell.We use count vectorizer and tf-idf vectorizer to vectorizer our model, preserving the terms occurring between 10% and 70% of the postings.As a result, each user post now has 595 features. The term frequency-inverse document frequency (tf-idf) is a metric that considers the significance of a word in relation to how often it appears in a document and corpus. To comprehend tf-idf, we must first comprehend term frequency and inverse document frequency.

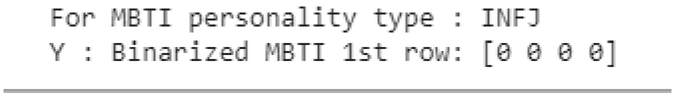
3.4.7 Splitting into X and Y variable

As a result, we've divided the characteristics into two categories:

X: TF-IDF representation of user posts

Y: Binaries MBTI form of personality type

Let's have a look at how the postings appear in TF-IDF: (For demonstrative purposes, we've taken the first post.)



```
For MBTI personality type : INFJ
Y : Binarized MBTI 1st row: [0 0 0 0]
```

Figure 3.22: personality Word cloud

We were able to transform the textual data into numerical representation effectively.

3.4.8 Training and Evaluating Models

RandomForestClassifier

XGBoost Model

SGDClassifier

Logistic Regression

KNN Model

SVM Model

3.5 Statistical Analysis

Table 3.1: Statistical Analysis

1. The total number of the columns are 2
2. Total number of rows 8675
3. 70% of the data we used in our model to train
4. 30% of the total data is used for testing purpose
5.Dataset is saved in csv file

CHAPTER 4

EXPERIMENTAL RESULTS, DISCUSSION AND CONCLUSION

4.1 Experimental Setup

In this section, we will talk about the experimental result. We've only gone through the first six algorithms we'll utilize. So now we'll find out how accurate those algorithms were. We'll also compare the accuracy of each of the six methods.

Steps that we follow to complete this research work are:

Step-1: Collect dataset

Step-2: Import different libraries

Step-3: Data visualization

Step-4: Pre-processing stage

Step-5: Feature Engineering

Step-6: Split our dataset

Step-7: Create models for all 6 algorithms

Step-8: Train with all the 6 different machine learning algorithms

Step-9: Find the accuracy of all the algorithms

These are the steps we follow to complete this research work

4.2 Experimental Result

We all know that no machine can provide us with 100% output. We may also train our model and tweak some parameters to improve accuracy and ensure correct training. However, the accuracy of the various algorithms is rather excellent. Below are some photographs that demonstrate our study effort concisely. We can see precision, recall, f1 score, support, and accuracy in these photographs.

Table 4.1: Results of Different Algorithm

Algorithm Name	**Introversion/ Extroversion (IE)**	**Intuition/ Sensing (NS)**	**Feeling/ Thinking (FT)**	**Judging/ Perceiving (JP)**
RandomForestClassifier	77.72	86.03	67.97	62.80
XGBoost Model	77.65	86.06	68.77	64.83
SGDClassifier	77.30	86.03	72.30	65.04
Logistic Regression	77.54	86.06	72.44	64.51
KNN Model	75.76	85.75	53.82	44.99
SVM Model	77.96	86.03	72.62	65.87

Table 4.2: Support Vector Machines results

Algoritm name	Precision		Recall		F1-score		Accuracy (%)
	0	1	0	1	0	1	
IE	0.78	0.80	1.00	0.01	0.88	0.01	0.78
NS	0.86	0.00	1.00	0.00	0.92	0.00	0.86
FT	0.74	0.71	0.77	0.67	0.75	0.69	0.73
J/P	0.63	0.67	0.31	0.88	0.42	0.76	0.66

4.3 Discussion

We can conclude our research work by saying that we have tried to use different algorithms to compare the accuracy of detecting Personality prediction. We get the highest accuracy from support vector machine or SVM. In this research paper, a few existing classification methods have been explained on the basis of accuracy for personality prediction. We create a classification method to show the accuracy of assessing a user's MBTI personality characteristics using only 50 tweets user.

CHAPTER 5

IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY

5.1 Impact on Society

The goal of our research is to increase business and advertise to customers by her personality. An organization can identify a person's personality in order to provide services, our research help to find out personality .it is very important to the modern generation. Anyone can find personality without meeting a specific person. So it is very significant to detect human behavior.

5.2 Impact on Environment

If anyone knows about a person's personality, they found everything about him. If they want to use this thing in a dishonest way, they can harm that person, data center has to make sure that it is used properly, then it will not affect the environment.

5.3 Ethical Aspects

According to the owner of the online business, we can advise customers in various ways about the product of their choice, and there are other companies as well. In the case of security, the prediction of the personality often gives an idea about the criminal. This is very important for the overall security of the country.

5.4 Sustainability Plan

The sustainable plan of this project is that through it everything about a person can be known directly or in a very short time. We can make any decision very quickly if we can find out fast a person's personality.

CHAPTER 6

SUMMARY, CONCLUSION AND IMPLICATION FOR FUTURE RESEARCH

6.1 Summary of the Study

We took the dataset from Kaggle in our research. Our dataset was a text dataset. So that firstly we preprocess our dataset. And this preprocessing dataset uses machine learning. We use some machine learning models to get the best accuracy. And using after 6 algorithms, we're getting the desired result. There has been a desired model through which we have got the most, and that is what we have used as the desired maximum accuracy.

6.2 Conclusions

We've been able to figure out human personality predictions through this project, and we've used a machine learning model to get people out of prediction. In this paper, we create a classification method to show the accuracy of assessing a user's MBTI personality characteristics using only 50 tweets. We use count-based vectorization (TF-IDF) and word embedding to incorporate language-based characteristics. Using six algorithm there are RandomForestClassifier, XGBoost Model, SGDClassifier, Logistic Regression, KNN Model, SVM Model.And the SVM model find out best accuracy in this project. Using the SVM model, It is possible to train the machine perfectly, and it will help the machine to understand more easily than others machine learning models.

6.3 Implication for Further Study

We are using some dataset, if we are using a huge dataset we can find many other people's personality predictions. This huge dataset uses online business, online marketing, etc.

REFERENCES

[1]. S. Bharadwaj, S. Sridhar, R. Choudhary and R. Srinath, "Persona Traits Identification based on Myers-Briggs Type Indicator(MBTI) - A Text Classification Approach," 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Bangalore, 2018, pp. 1076-1082.

[2] M. Gjurković and J. Šnajder, “Reddit: A Gold Mine for Personality Prediction,” In Proceedings of the Second Workshop on Computational Modeling of People’s Opinions, Personality, and Emotions in Social Media , pp. 87-97, 2018.

[3] 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), Yogyakarta, 2015, pp. 170-174.

[4] M. Gjurković and J. Šnajder, “Reddit: A Gold Mine for Personality Prediction,” In Proceedings of the Second Workshop on Computational Modeling of People’s Opinions, Personality, and Emotions in Social Media , pp. 87-97, 2018.

[5] B. Plank, and D. Hovy, “Personality traits on twitter—or—how to get 1,500 personality tests in a week.” In Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pp. 92-98, 2015.

[6] F. Celli and L. Rossi, “The role of emotional stability in Twitter conversations,” In Proceedings of the workshop on semantic analysis in social media, Association for Computational Linguistics, pp. 10-17, 2012.

[7] F. Celli, “Unsupervised personality recognition for social network sites,” In Proc. of Sixth International Conference on Digital Society, 2012.

[8] N. R. Ngatirin, Z. Zainol and T. L. C. Yoong, "A comparative study of different classifiers for automatic personality prediction," 2016 6th IEEE International Conference on Control System, Computing and Engineering (ICCSCE), Batu Ferringhi, 2016, pp. 435-440.

[9] 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), Yogyakarta, 2015, pp. 170-174.

[10] D. Nielsen, “Tree Boosting With XGBoost-Why Does XGBoost Win Every Machine Learning Competition? (Master's thesis, NTNU),” 2016.

[11] M. M. Tadesse, H. Lin, B. Xu and L. Yang, "Personality Predictions Based on User Behavior on the Facebook Social Media Platform," in IEEE Access, vol. 6, pp. 61959-61969, 2018.

[12] D. Azucar, D. Marengo, and M. Settanni. Predicting the big 5 personality traits from digital footprints on social media: A meta-analysis. Personality and Individual Differences, 124:150 – 159, 2018

[13] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M. E. Seligman, et al. Personality, gender, and age in the language of social media: The open-vocabulary approach. PloS one, 8(9):e73791, 2013.

[14] N. Alsadhan and D. Skillicorn, "Estimating Personality from Social Media Posts," 2017 IEEE International Conference on Data Mining Workshops (ICDMW), New Orleans, LA, 2017, pp. 350-356.

[15]M. C. Komisin and C. I. Guinn, “Identifying personality types using document classification methods,” In Twenty-Fifth International FLAIRS Conference, 2012.

[16]A. Tripathy, A. Agrawal and S. K. Rath, “Classification of sentiment reviews using n-gram machine learning approach,” Expert Systems with Applications, 57, pp. 117-126, 2016.

[17]M. Z. Asghar, A. Khan, F. Khan and F. M. Kundi, “RIFT: A Rule Induction Framework for Twitter Sentiment Analysis,” Arabian Journal for Science and Engineering, vol. 43, no. 2, pp.857-877, 2018.

[18]L. C. Lukito, A. Erwin, J. Purnama and W. Danoekoesoemo, "Social media user personality classification using computational linguistic," 2016 8th International Conference on Information Technology and Electrical Engineering (ICITEE), Yogyakarta, 2016, pp. 1-6

[19]M. Arroju, A. Hassan, and G. Farnadi, “Age, gender and personality recognition using tweets in a multilingual setting,” In 6th Conference and Labs of the Evaluation Forum (CLEF 2015): Experimental IR meets multilinguality, multimodality, and interaction, pp. 23-31, 2015

[20] J. Golbeck, C. Robles, M. Edmondson and K. Turner, "Predicting Personality from Twitter," 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing, Boston, MA, 2011, pp. 149-156.

[21]V. Ong, A. D. Rahmanto, Williem and D. Suhartono, “Exploring Personality Prediction from Text on Social Media: A Literature Review,” INTERNETWORKING INDONESIA, vol. 9, no. 1, pp. 65- 70, 2017a

[22]P. Arnoux, A. Xu, N. Boyette, J. Mahmud, R. Akkiraju, and V. Sinha. 25 tweets to know you: A new model to predict personality with social media. Computing Research Repository, arXiv, 2017.

[23]B. Plank and D. Hovy. Personality traits on twitter –or– how to get 1,500 personality tests in a week. In Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 92–98, 2015

[24]D. Quercia, M. Kosinski, D. Stillwell, and J. Crowcroft. Our twitter profiles, our selves: Predicting personality with twitter. In 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust, pages 180–185, Oct 2011.

[25]M. Sultana, P. P. Paul, and M. Gavrilova. Social behavioral biometrics: An emerging trend. International Journal of Pattern Recognition and Artificial Intelligence, 29(08):1556013, 2015.

[26]G. Seidman. Self-presentation and belonging on facebook: How personality influences social media use and motivations. Personality and Individual Differences, 54(3):402– 407, 2013

[27]B. W. Roberts, N. R. Kuncel, R. Shiner, A. Caspi, and L. R. Goldberg. The power of personality: The comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. Perspectives on Psychological Science, 2(4):313–345, 2007

[28]A. Furnham. The big five versus the big four: the relationship between the myers-briggs type indicator (mbti) and neo-pi five factor model of personality. Personality and Individual Differences, 21(2):303 – 307, 1996

[29]T. L. C. Yoong, N. R. Ngatirin, and Z. Zainol, "Personality prediction based on social media using decision tree algorithm," Pertanika J. Sci. Technol., vol. 25, no. S4, pp. 237–248, 2017

[30]N. R. Ngatirin, Z. Zainol, and T. L. C. Yoong, "A comparative study of different classifiers for automatic personality prediction," Proc. - 6th IEEE Int. Conf. Control Syst. Comput. Eng. ICCSCE 2016, pp. 435– 440, 2017.

[31]T. Tandera, Hendro, D. Suhartono, R. Wongso, and Y. L. Prasetio, "Personality Prediction System from Facebook Users," Procedia Comput. Sci., vol. 116, pp. 604–611, 2017

[32]W. Rc, Y. Munas, K. Cs, F. Ta, and Vithana N, "Personality Based ERecruitment System," Int. J. Innov. Res. Comput. Commun. Eng., vol. 5, 2017

[33]T. Tandera, Hendro, D. Suhartono, R. Wongso, and Y. L. Prasetio, "Personality Prediction System from Facebook Users," Procedia Comput. Sci., vol. 116, pp. 604–611, 2017.

[34]N. R. Ngatirin, Z. Zainol, and T. L. C. Yoong, "A comparative study of different classifiers for automatic personality prediction," Proc. - 6th IEEE Int. Conf. Control Syst. Comput. Eng. ICCSCE 2016, pp. 435– 440, 2017.

[35] Mayuiri Pundlik Kalghatgi, Mannjuli Rammanavar, Dr. Nandini Sindal "A Neural Network Approach to Personality Prediction based on the BigFive Model" in International Journal of Innovative Research in Advanced Engineering, Issue 8, Volume 2, August 2015.

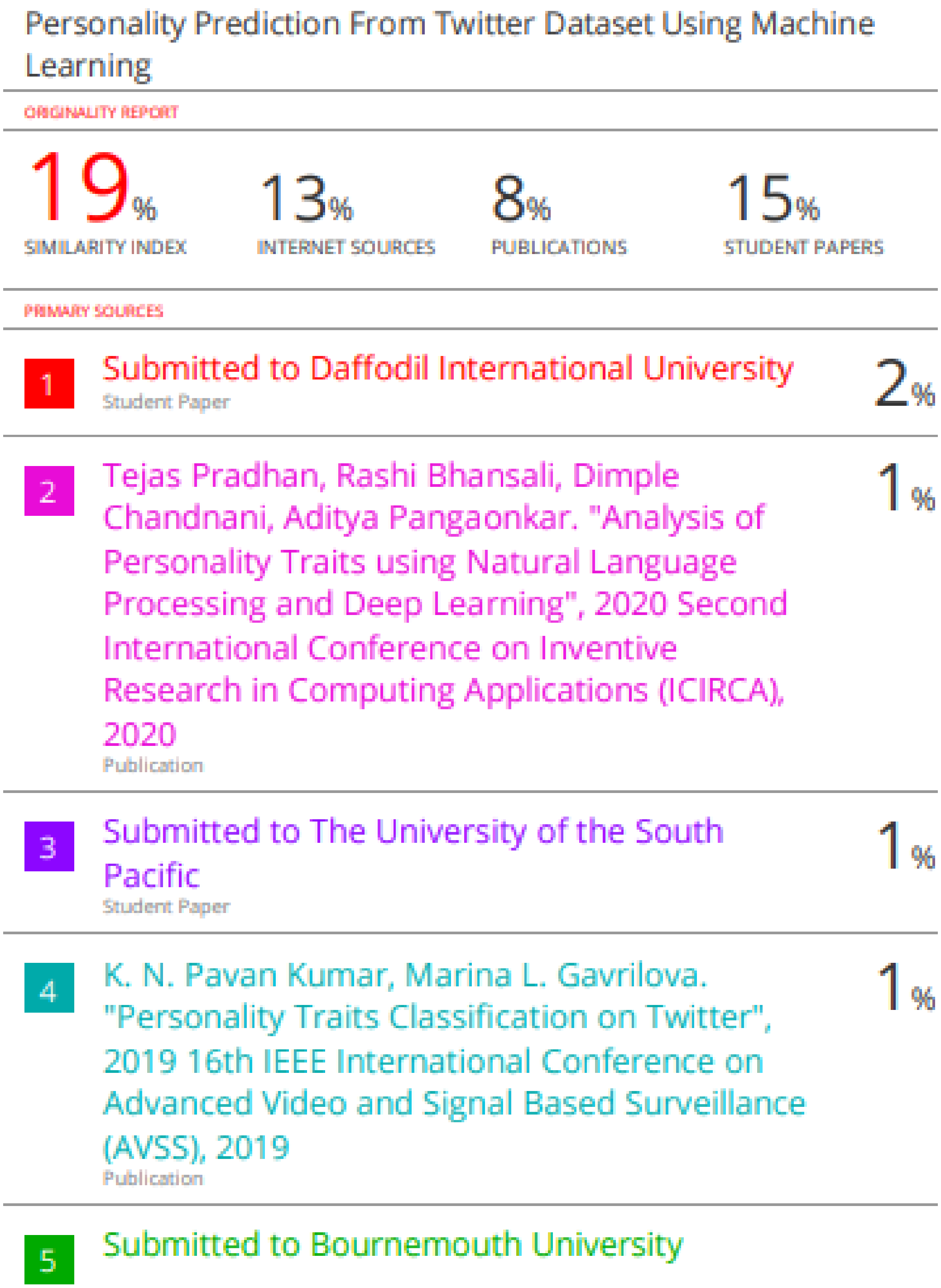
[36]Navonil Majumder, Soujanya Poria, Alexander Gelbukh, Erik Cambria, "Deep learning-based document modeling for personality detection from text," published by the IEEE Computer Society , IEEE Intelligent Systems 2017.

[37]Manasi Ombhase, Prajakta Gogate, Tejas Patil, Karan Nair and Prof. Gayatri Hegde, "Automated Personality Classification using Data Mining Techniques" Pillai Institute of Information Technology.

[38] Joel Philip, "Machine Learning for Personality Analysis Based on Big Five Model," in IC Joel Philip, "Machine Learning for Personality Analysis Based on Big Five Model," in ICDMAI, Volume 2 in January 2019.

[39]Jayashree Rout, Sudhir Bagade, Pooja Yede, Nirmiti Patil, "Personality evaluation and CV analysis using machine learning algorithm" in IJCSE, May 2019.

[40]Clemens Stachl, Florian Pargent Sven, Hilbert Gabriella M. Harari, Ramona Schoedel, Sumer Vaid, Samuel D. Gosling, Markus Buhner, "Personality Research & Assesment in the era of machine learning" in European Journal of Personality, 28 May 2020.



Personality Prediction From Twitter Dataset Using Machine Learning

ORIGINALITY REPORT

19%	13%	8%	15%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	2%
2	Tejas Pradhan, Rashi Bhansali, Dimple Chandnani, Aditya Pangaonkar. "Analysis of Personality Traits using Natural Language Processing and Deep Learning", 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), 2020 Publication	1%
3	Submitted to The University of the South Pacific Student Paper	1%
4	K. N. Pavan Kumar, Marina L. Gavrilova. "Personality Traits Classification on Twitter", 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2019 Publication	1%
5	Submitted to Bournemouth University	