

**PERSONALITY PREDICTION USING VARIOUS MACHINE LEARNING  
ALGORITHMS**

A DISSERTATION

SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE  
AWARD OF DEGREE  
OF  
MASTER OF TECHNOLOGY  
IN  
**INFORMATION SYSTEMS**

Submitted by:

**NIRMAL ARYAL**  
**2K21/ISY/17**

Under the supervision of  
**Dr. Ritu Agarwal**



**DEPARTMENT OF INFORMATION TECHNOLOGY**  
**DELHI TECHNOLOGICAL UNIVERSITY**  
(Formerly Delhi College of Engineering)  
Bawana Road, Delhi-110042

JUNE, 2023

**DEPARTMENT OF INFORMATION TECHNOLOGY**

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi – 110042

**CANDIDATE'S DECLARATION**

I, Nirmal Aryal Roll No. 2K21/ISY/17 student of M. Tech (Information Systems), hereby declare that the project Dissertation titled “PERSONALITY PREDICTION USING VARIOUS MACHINE LEARNING ALGORITHMS” which is submitted by me to the Department of Computer Science & Engineering, Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of the degree of Master of Technology, is original and not copied from any source without proper citation. This work has not previously formed the basis for the award of and Degree, Diploma Associateship, Fellowship or other similar title or recognition.

Place: Delhi

NIRMAL ARYAL

Date:

2K21/ISY/17

**DEPARTMENT OF INFORMATION TECHNOLOGY**

**DELHI TECHNOLOGICAL UNIVERSITY**

(Formerly Delhi College of Engineering)

Bawana Road, Delhi – 110042

**CERTIFICATE**

I hereby certify that the Project Dissertation titled “**PERSONALITY PREDICTION USING VARIOUS MACHINE LEARNING ALGORITHMS**” which is submitted by Nirmal Aryal, Roll No. 2K21/ISY/17, Information Systems, Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of the degree of Master of Technology, is a record of the project work carried out by the students under my supervision. To the best of my knowledge, this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Place: Delhi

**DR. RITU AGARWAL**

Date:

**SUPERVISOR**

## **DEPARTMENT OF INFORMATION TECHNOLOGY**

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi – 110042

### **ACKNOWLEDGMENT**

The success of this project depends on the help and contribution of a large number of people as well as the organization. I am grateful to everyone who contributed to the project's success.

I'd want to convey my gratitude to **DR. RITU AGARWAL**, my project guide, for allowing me to work on this research under his supervision. His unwavering support and encouragement have taught me that the process of learning is more important than the ultimate result. Throughout all of the progress reviews, I am appreciative to the panel faculty for their assistance, ongoing monitoring, and motivation to complete my project. They assisted me with fresh ideas, gave crucial information, and motivated me to finish the task.

NIRMAL ARYAL

2K21/ISY/17

## ABSTRACT

The relevance of personality prediction and its possible advantages are examined in this thesis study. Since it can help us better understand human behaviour and individual characteristics, personality prediction is vital. The study highlights the value of psychological knowledge gathered through personality prediction in variety of industries, including marketing, education, staffing, mental health, and personal growth. Personalized marketing efforts, targeted product suggestions, and specialised learning tactics are all made possible by personality prediction since it allows experiences and services to be tailored to individual preferences. It helps with educated recruiting decisions and improved job fit in the area of human resources. Furthermore, early detection of personality features linked to mental health issues enables prompt assistance and intervention.

Using assessment measures, the study compares the effectiveness of different algorithms. The use of several machine learning and deep learning algorithms allows for a thorough examination of personality prediction methods. The algorithms taken into account in this thesis were chosen to reflect a wide range of methodologies, including ensemble techniques, conventional machine learning, and deep learning models. By comparing the outcomes of these algorithms using evaluation criteria, their advantages and disadvantages in personality prediction tasks are highlighted.

The research's findings offer insightful information on how various personality prediction algorithms work. This information can assist in the creation of more precise and successful personality prediction models, enabling the use of personality prediction in real-world contexts. The thesis also advances the science of machine learning by demonstrating the strengths and weaknesses of several algorithms in the context of personality prediction. The ethical issues around privacy, accuracy, and appropriate use of personal information in personality prediction must be taken into account, though. This research increases our knowledge of human behaviour by addressing these issues and utilising the potential advantages of personality prediction. It also prepares the way for future improvements in tailored experiences, hiring procedures, mental health assistance, and personal growth.

# CONTENTS

|   |             |
|---|-------------|
| <b>Candidate's Declaration</b>                            | <b>i</b>    |
| <b>Certificate</b>  | <b>ii</b>   |
| <b>Acknowledgement</b>                                    | <b>iii</b>  |
| <b>Abstract</b>   | <b>iv</b>   |
| <b>Contents</b>   | <b>v</b>    |
| <b>List of Tables</b>                                     | <b>vii</b>  |
| <b>List of Figures</b>                                    | <b>viii</b> |
| <b>List of Abbreviations</b>                              | <b>ix</b>   |
| <b>CHAPTER 1 INTRODUCTION</b>                             | <b>1</b>    |
| <b>CHAPTER 2 PRIOR WORK</b>                               | <b>4</b>    |
| <b>CHAPTER 3 FUNDAMENTALS</b>                             | <b>11</b>   |
| 3.1 Text based classification                             | 11          |
| 3.2 Natural Language Processing                           | 13          |
| 3.3 Machine Learning                                      | 14          |
| 3.3.1 Supervised Learning                                 | 15          |
| 3.3.2 Unsupervised Learning                               | 16          |
| <b>CHAPTER 4 DATASET DESCRIPTION</b>                      | <b>17</b>   |
| <b>CHAPTER 5 PROPOSED WORK</b>                            | <b>24</b>   |
| 5.1 Problem statement                                     | 24          |
| 5.2 Proposed Methodology                                  | 25          |
| 5.2.1 Data Extraction                                     | 26          |
| 5.2.2 Machine Learning models                             | 28          |
| <b>CHAPTER 6 EXPERIMENTS AND RESULTS</b>                  | <b>34</b>   |
| 6.1 Analysis and Visualization of the Experimental Result | 34          |

|  |           |
|--|-----------|
| 6.2 Stacked LSTM Model                       | 36        |
| 6.3 RNN Model                                | 38        |
| 6.4 BERT Model                               | 40        |
| <b>CHAPTER 7 CONCLUSION AND FUTURE SCOPE</b> | <b>41</b> |
| 7.1 Conclusion                               | 41        |
| 7.2 Future scope and recommendations         | 41        |
| <b>REFERENCES</b>                            | <b>43</b> |

# LIST OF TABLES

- 2.1: Summarized review of literature papers
- 6.1: Illustration of Stacked LSTM model architecture
- 6.2: Illustration of RNN model architecture



# LIST OF FIGURES

- 3.1: Flow of text classification
- 3.2: Some areas where NLP is used
- 3.3: Machine learning Applications
- 3.4: Workflow of supervised learning algorithms
- 3.5: Workflow of Unsupervised learning algorithms
- 4.1: Snapshot of Dataset
- 5.2: Snapshot of list of stop words
- 5.3: Figure illustrating the working of logistic regression algorithm for personality prediction
- 5.4: Figure illustrating the working of logistic regression algorithm for personality prediction
- 6.1: Comparison of machine learning classifiers
- 6.2: Flow of Stacked LSTM model
- 6.3: Flow of RNN model
- 6.4: Execution of BERT model algorithm

# LIST OF ABBREVIATIONS

|               |   |
|---------------|---|
| <b>NLP</b>    | Natural Language Processing               |
| <b>SNSs</b>   | Social Networking Sites                   |
| <b>ML</b>     | Machine Learning                          |
| <b>SVM</b>    | Support Vector Machine                    |
| <b>CNN</b>    | Convolutional Neural Network              |
| <b>HIE</b>    | Heterogenous Information ensemble         |
| <b>BOW</b>    | Bag of Words                              |
| <b>TF-IDF</b> | Term Frequency-Inverse Document Frequency |

# CHAPTER 1

## INTRODUCTION

A key component of human behaviour, personality affects how people view the environment, interact with others, and make decisions. Numerous fields, including marketing, education, recruiting, and mental health, can benefit significantly by understanding personality characteristics and their consequences. Historically, self-report questionnaires like the Myers-Briggs Type Indicator (MBTI) have been used for personality assessment. However, because people need to answer a lot of questions, these examinations might take quite a long time.

This thesis seeks to build a different method of personality type prediction using Reddit data in order to overcome this problem. We want to develop a model that properly categorises people into certain personality types, such as INTP, INTJ, ENTJ, ENTP, and INFJ, by utilising the enormous quantity of information supplied by users on our internet platform.

The demand for a quicker and more accurate form of personality assessment is what drove this research. The extensive questionnaire for the MBTI test can be tiresome and time-consuming for people, despite its widespread use. Reddit data may be used to examine linguistic patterns, debates, posts, and comments to learn more about people's personalities without requiring them to explicitly respond to queries. Reddit data may be used to predict personalities in a number of ways. First off, it offers a huge and varied dataset that includes debates on a variety of topics. This variation enables a thorough examination of linguistic and behavioural patterns across several

personality types. We can understand the subtleties and complexities of this data by mining it.

Secondly, the model created utilising Reddit data may make predictions in real-time, increasing the effectiveness and usability of personality evaluation. The programme can quickly analyse massive amounts of data and offer immediate findings rather than waiting for people to finish drawn-out surveys and human grading procedures. The model is extremely useful for applications that call for short personality assessments, such as hiring procedures or personal development initiatives, due to its speed and efficiency. Using cutting-edge machine learning algorithms, precise personality predictions will be made. The model will go through rigorous training using labelled Reddit data, which contains information on users' self-reported personality types. Random Forest, Decision Tree, Logistic Regression, Gradient Boosting, Ada Boosting, Recurrent Neural Network (RNN) with 2-GRU layer, K-Nearest Neighbors (KNN), Stacked Long Short-Term Memory (LSTM), and BERT are just a few of the classification algorithms that will be used during this training process. These algorithms provide several methods for identifying patterns in textual material and can be used to pinpoint the characteristics that set distinct personality types apart. Comparing the model's predictions to the self-reported personality types found in the MBTI 500 dataset will be used to assess the model's performance. This dataset is a compilation of answers from people who have completed the MBTI questionnaire. We may evaluate the model's performance in predicting personality types by using assessment measures including accuracy, precision, recall, and F1 score. This examination will offer perceptions into the model's strengths and weaknesses and guide further improvements in its performance.

The results of this study have broad applications in many different sectors. The model can enable tailored experiences in marketing, education, and counselling by providing a quicker and more effective form of personality prediction. It can help with developing specialised learning tactics, making focused product suggestions, and locating job pathways that fit people's personality types. The approach can speed up candidate evaluations during hiring procedures and enhance the fit between applicants and job specifications. Additionally, the model may have an influence on mental health and personal growth. Mental health providers can get important insights into people's

vulnerabilities and customise therapies by correctly predicting personality types. Individuals may also gain from self-awareness and personal growth by comprehending their personality features and using that information to make better decisions for themselves.

In essence, the goal of this thesis is to create a model that predicts personality types using data from Reddit, providing a quicker and more effective substitute for the conventional MBTI exam. The model is able to catch subtleties in people's language use and behaviour since Reddit data is a rich, varied, and plentiful source of information. This study intends to enhance the science of personality prediction and its useful applications in a variety of fields by utilising cutting-edge machine learning techniques and assessing the model's performance against the MBTI 500 dataset.

## CHAPTER 2

### PRIOR WORK

The criterion validity of lower-level Conscientiousness qualities is examined by (O'Neill et al., 2011) to ascertain whether any one trait is particularly predictive of team performance. The authors of (Skowron et al., 2016) suggest an unique technique that combines text, picture, and user meta information from Twitter and Instagram, two separate SNSs. On a wide range of social media ground truth data from Facebook, Twitter, and YouTube, (Farnadi et al., 2016) undertake a comparative examination of cutting-edge computational personality assessment approaches. More specifically, (Salam et al., 2017) explore how participants' personalities might be combined with the personality of the robot to predict each participant's level of interest. In order to predict users' personality traits by integrating heterogeneous information, such as self-language, avatar, emoticon, and responsive patterns (Wei et al., 2017) offer a Heterogeneous Information Ensemble framework, known as HIE. Other influential work includes (Paunonen et. al., 1985), (Sigvardsson et. al., 1987), (Ozer et. al., 2006), (Lima et. al., 2014), (Kluemper et. al., 2014).

(Kampman et al., 2018) offer a tri-modal architecture with distinct channels for voice, text, and video data to predict Big Five personality trait scores from video snippets. (Hinds et al., 2019) respond to three inquiries: What is presently known regarding computer- and human-based personality evaluations (Hinds et al., 2019)? (Ro et al., 2019) suggest creating distinct prediction models for each dimension, looking at the function of sociodemographic and personality factors (such as the "Big Five" and the "Dark Tetrad"). The goal of (Cai et al., 2019) was to use connectome-based predictive modelling (CPM), a recently established machine learning method, to build a trustworthy prediction model of personality in a large sample. The Pearson

correlation is one of five feature selection techniques that (Marouf et al., 2020) compares for performance with others namely correlation-based feature subset (CFS), information gain (IG), symmetric uncertainty (SU) evaluator, and chi-squared (CHI) method. (Lynn et. al., 2020) present a novel model that uses message-level attention to learn the relative weight of users' social media posts for assessing their five factor personality traits.

The focus has been mostly on individual personality traits, mainly ignoring relationship aspects of personality, despite a noticeable increase in research on personality prediction based on online activity. By offering a prediction model for a comprehensive personality profiling in OSNs that incorporated socio-relational qualities (attachment orientations) in addition to typical personality traits, (Karanatsiou et al., 2020) seek to close this gap. (Revelle et al., 2021) examine data collecting and analysis techniques that place an emphasis on how well-suited specific personality traits are for forecasting real-world standards (e.g., smoking, exercise, self-rated health). Digital twins in human understanding: a deep learning-based approach to character recognition is the subject of a research by (Sun et al., 2021). By analysing users' posts, a novel method to formalise personality as digital twin models is offered. Other influential work includes (Zhao et. al., 2020).

In order to do this, (Ramezani et al., 2022) offer the Knowledge Graph Attention Network text classifier KGrAt-Net. The Big Five personality qualities are the foundation of the unique knowledge graph-enabled method to text-based APP presented by (Ramezani et al., 2022). Functional connectivity of the central autonomic and default mode networks reflects neural correlates and predictors of unique personality, according to a study by (Li et al., 2022). The five-factor model was used to evaluate personality traits. The creation of a socio-technical framework for evaluating the stability of algorithmic systems is the focus of (Rhea et al., 2022). (Marrero et al., 2022) investigate the potential predictive value of linguistic and vocal information derived from semi-structured voice samples for traditional personality assessments.

The summary of the complete literature research, combined with their advantages and shortcomings, are stated in Table 2.1.

Table 2.1: Summarized review of literature papers

| REFERENCE | METHODOLOGY  | ADVANTAGES  | LIMITATIONS  |
|-----------|--|---|--|
| [1]       | Proposes KGrAt-Net, a Knowledge Graph Attention Network text classifier  | - Utilizes Knowledge Graph Attention Network for text classification      | - Specific details about the methodology and performance of KGrAt-Net are not provided |
| [2]       | Presents a knowledge graph-enabled approach to text-based APP based on the Big Five personality traits   | - Integrates knowledge graph into text-based APP for personality traits   | - Specific details about the knowledge graph-enabled approach are not provided         |
| [3]       | Studies functional connectivity of the central autonomic and default mode networks as neural correlates and predictors of individual personality | - Examines the neural correlates and predictors of individual personality | - Specific details about the study's methodology and findings are not provided         |
| [4]       | Develops a socio-technical framework for auditing the stability of algorithmic systems   | - Provides a framework for auditing the stability of                      | - Specific details about the socio-technical framework are not provided                |



|     |   | algorithmic systems  |  |
|-----|---|--|--|
| [5] | Examines the use of linguistic and vocal information from semi-structured vocal samples to predict conventional measures of personality | - Investigates the potential of linguistic and vocal information for personality prediction            | - Specific details about the methodology and results are not provided                              |
| [6] | Constructs a benchmark called Story2Personality to support the development of neural models for personality prediction                  | - Provides a benchmark dataset for training and evaluating neural models in personality prediction     | - Specific details about the Story2Personality benchmark are not provided                          |
| [7] | Demonstrates how textual content from answers to interview questions can be used to infer personality traits                            | - Explores the use of textual content from interview answers for personality trait inference           | - Specific details about the methodology and performance of personality inference are not provided |
| [8] | Develops and validates a machine learning model with domain knowledge introduced to enhance accuracy                                    | - Integrates domain knowledge into a machine learning model for enhanced accuracy and interpretability | - Specific details about the methodology and validation results are not provided                   |

|      |  |  |   |
|------|--|--|---|
|      | and improve interpretability   |  |   |
| [9]  | Proposes a two-step approach to map sentences based on hierarchical memberships and polarity   | - Introduces a two-step approach for sentence mapping based on hierarchical memberships and polarity | - Specific details about the two-step approach and its performance are not provided |
| [10] | Integrates text, image, and users' meta features from Twitter and Instagram to develop a novel method  | - Utilizes multiple data sources (text, image, meta features)  | - Does not provide specific details about the integrated method                     |
| [11] | Performs a comparative analysis of state-of-the-art computational personality recognition methods on social media data from Facebook, Twitter, and YouTube | - Provides a comparative analysis of personality recognition methods                                 | - Does not provide specific details about the comparative analysis                  |
| [12] | Investigates how participants' personalities can be used together with   | - Examines the role of participants' and robot's personalities in                                    | - Specific details about the methodology are not provided                           |

|      |   |   |   |
|------|---|---|---|
|      | the robot's personality to predict the engagement state of each participant   | predicting engagement state   |   |
| [13] | Proposes a Heterogeneous Information Ensemble framework called HIE to predict users' personality traits by integrating self-language usage, avatar, emoticon, and responsive patterns | - Integrates heterogeneous information to predict users' personality traits | - Does not provide specific details about the HIE framework |
| [14] | Influential work on personality traits  | - Contributes to the understanding of personality traits                    | - Specific details about the study are not provided         |
| [15] | Influential work on personality traits  | - Contributes to the understanding of personality traits                    | - Specific details about the study are not provided         |
| [16] | Influential work on personality traits  | - Contributes to the understanding of personality traits                    | - Specific details about the study are not provided         |
| [17] | Influential work on personality traits  | - Contributes to the understanding of personality traits                    | - Specific details about the study are not provided         |

|      |  |   |   |
|------|--|---|---|
| [18] | Influential work on personality traits   | - Contributes to the understanding of personality traits  | - Specific details about the study are not provided                         |
| [19] | Proposes a tri-modal architecture using audio, text, and video data from video clips to predict Big Five personality trait scores          | - Utilizes multiple modalities (audio, text, video) for predicting personality traits               | - Specific details about the tri-modal architecture are not provided        |
| [20] | Examines human- and computer-based personality assessments and addresses questions related to the current understanding of the assessments | - Provides insights into the current knowledge of human- and computer-based personality assessments | - Specific details about the methodology used in the study are not provided |

# CHAPTER 3

## FUNDAMENTALS

### 3.1 Text based classification

Text-based classification, sometimes referred to as text classification or text categorization, is a key activity in natural language processing (NLP) that entails automatically classifying or labelling text documents according to predetermined categories or labels. In many applications, including sentiment analysis, spam detection, subject categorization, document organisation, and many more, it is essential.

Creating models that can properly and effectively categorise vast amounts of textual data is the aim of text-based classification. The first phase in this procedure is data pre-processing, which entails cleaning, tokenizing, and converting the text into numerical representations appropriate for machine learning algorithms. The essential information is subsequently extracted from the text using feature extraction techniques like bag-of-words (BOW), term frequency-inverse document frequency (TF-IDF), or word embeddings.

Text categorization problems are frequently carried out using machine learning techniques like Naive Bayes, Support Vector Machines (SVM), or more recently, deep learning models like Convolutional Neural Networks (CNN) and Transformer-based models. These models discover patterns and connections between the input text and relevant categories by learning from labelled training data. By automatically learning hierarchical representations

of the text input, they are able to manage the complexity and diversity of human language.

Metrics that measure the effectiveness of the model in accurately categorising the text documents, such as accuracy, precision, recall, and F1 score, are frequently used in the evaluation of text-based classification models [30]. The choice of features, the choice of methods, and the quality and representativeness of the training data all have a significant impact on how well these models function. Cross-validation and hyperparameter tweaking are two methods that are frequently used to make sure that the models are resilient and generalizable.

Text-based categorization has many uses across many industries. In customer feedback analysis, it lets companies to automatically analyse and classify significant amounts of customer reviews, social media postings, or survey results, assisting them in gaining insightful knowledge on the attitudes and preferences of their customers. It makes textual data structure and retrieval easier in document organisation, enabling effective content management and information retrieval. It may be used in content moderation to recognise and filter offensive or dangerous information, creating a safer and more welcoming online environment.

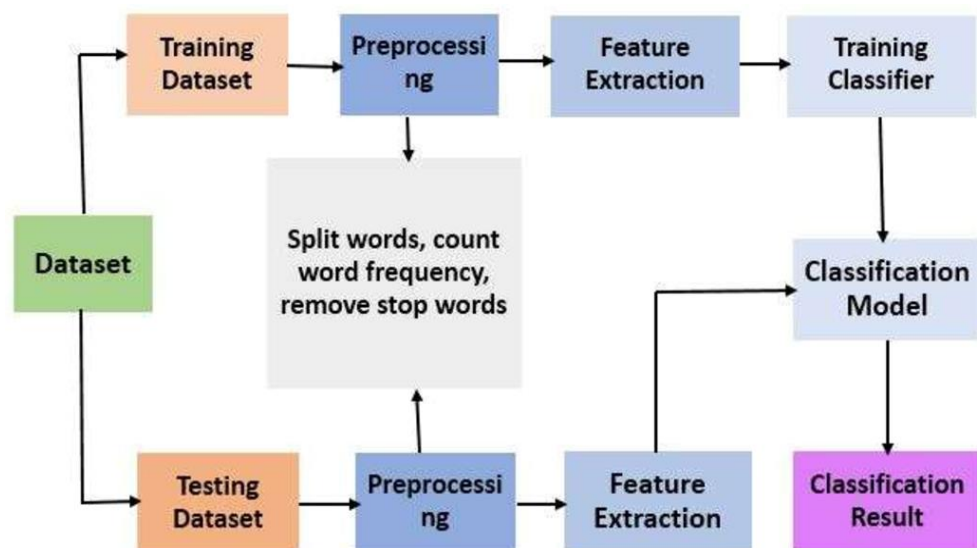


Fig. 3.1: Flow of text classification [21]

### 3.2 Natural Language Processing

How computers and human language interact is the focus of the field of study called "Natural Language Processing" (NLP). Its fundamental goal is to enable computers to understand, translate, and compose human language.

NLP includes several phases, including text preparation, feature extraction, and modelling. The text pre-processing stage involves cleaning up and converting the raw text input into a format that NLP algorithms can comprehend. Stop words like "the," "an," and "a" can be removed using tokenizing, stemming (the process of breaking down words into their most fundamental components), and other methods.

Relevant textual characteristics are found and retrieved during the feature extraction stage. The existence of specific grammatical structures, the volume of the text, or the frequency of particular words or phrases are a few examples. Then, machine learning models that can automatically categorise or produce text are trained using these attributes.

Numerous applications, such as social media monitoring, customer feedback analysis, and product review analysis, all employ sentiment analysis. NLP algorithms must be able to comprehend both the meaning of words and the context in which they are being used in order to perform sentiment analysis.

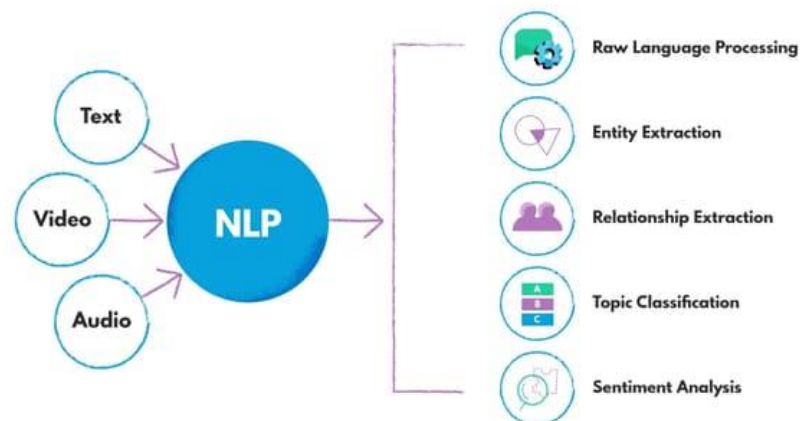


Fig. 3.2: Some areas where NLP is used. [22]

### 3.3 Machine Learning

Artificial intelligence (AI) has a subset called machine learning (ML) that focuses on creating algorithms and models that let computers learn and make predictions or judgments without having to be explicitly programmed. It is predicated on the notion that computers are capable of learning from data, seeing patterns, and formulating wise judgments or predictions.

The fundamental idea behind machine learning is to train a model with labelled data so that it may learn from the correlations and patterns found in the data. The model may then apply this learning to anticipate the future or make judgments based on brand-new, unexplored data. Training and inference are the two key steps in this process [33].

Once trained, the model may be used to draw conclusions from fresh, unexplored data. Based on the patterns it has discovered while studying the training set of data, the model uses the input attributes to provide predictions or judgments. This gives the model the ability to categorise objects, forecast numerical values, identify abnormalities, provide product recommendations, or carry out other activities depending on the particular problem it was trained for.

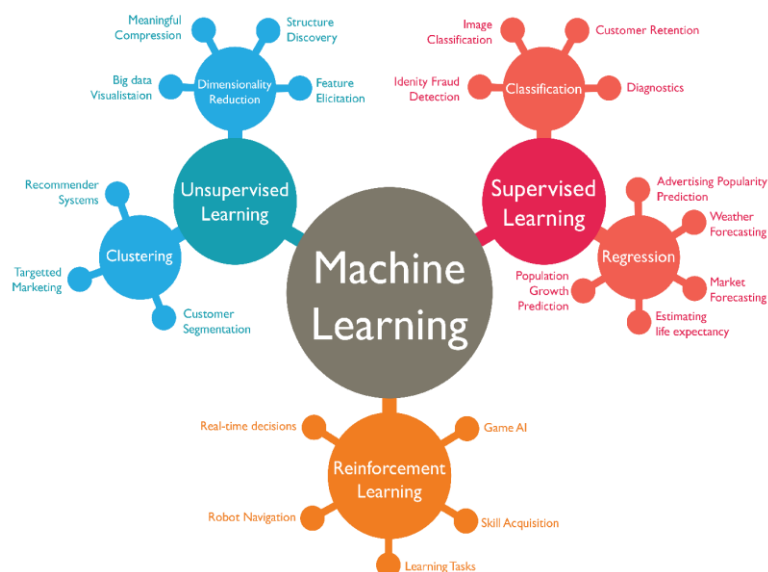


Fig. 3.3: Machine learning Applications. [23]



### 3.3.1 Supervised Learning

A key idea in machine learning is supervised learning, where a model learns from labelled training data to make judgments or predictions on brand-new, unlabeled data. In supervised learning, input features and related target labels make up the training data. By examining the patterns and correlations in the data, the model learns to map the input attributes to the appropriate labels.

The labelled data is fed into the model during training, and its internal parameters are changed depending on the discrepancies between the projected outputs and the actual labels [35]. The goal is to maximise the model's performance by reducing the error between the anticipated and actual values.

When a model has been trained, it may be applied to infer new data by taking characteristics from the input and producing predictions or judgments based on the patterns it has discovered. In several tasks, including classification, regression, and object identification, supervised learning is applied.

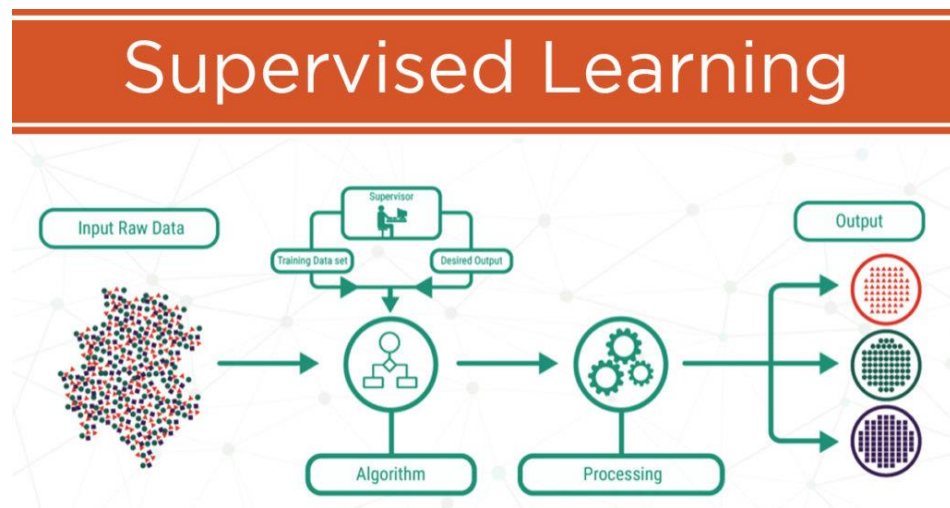


Fig. 3.4: Workflow of supervised learning algorithms. [24]

### 3.3.2 Unsupervised Learning

The goal of unsupervised learning, a subfield of machine learning, is to find structures, correlations, and patterns in unlabeled data. Unsupervised learning does not require labelled examples or predetermined goal outputs, in contrast to supervised learning. Instead, it seeks to draw insights directly from the data.

Clustering algorithms group similar data points together based on their inherent similarities, enabling the identification of natural clusters within the data. Dimensionality reduction techniques aim to reduce the complexity of high-dimensional data by extracting a lower-dimensional representation while preserving important features. Anomaly detection focuses on identifying unusual or abnormal data points that deviate from the norm.

Unsupervised learning is useful for large-scale, unstructured data pre-processing, exploratory data analysis, and insight generation. It may be used in many different contexts, such as cybersecurity anomaly detection, consumer segmentation, picture and text analysis, and recommendation systems.

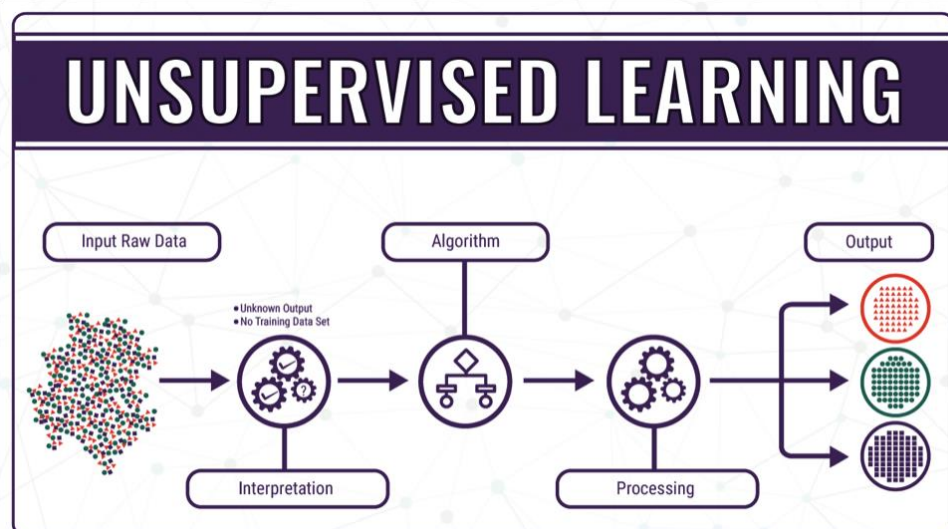


Fig. 3.5: Workflow of Unsupervised learning algorithms

# CHAPTER 4

## DATASET DESCRIPTION

A popular introspective self-report questionnaire, the MBTI (Myers-Briggs Type Indicator) tries to evaluate and categorise various psychological preferences in how people view the world and make decisions. The MBTI was created by Katharine Cook Briggs and Isabel Briggs Myers and is based on Carl Jung's idea of psychological types.

Individuals are divided into four pairings according to their psychological preferences in the MBTI framework, which produces sixteen different personality types. Extraversion (E) or Introversion (I), Sensing (S) or Intuition (N), Thinking (T) or Feeling (F), and Judging (J) or Perceiving (P) are some of these inclinations. The MBTI offers insights into a person's communication style, problem-solving methodology, decision-making inclinations, and interpersonal dynamics by analysing their preferences across various domains.

The MBTI 500 dataset, which is a useful tool for studying personality characteristics and examining the connection between psychological preferences and different outcomes, is the main subject of the thesis research. The results from people who have completed the MBTI questionnaire are included in the MBTI 500 dataset. Both demographic information and data on people's self-reported personality types are included.

The dataset enables statistical analysis and investigation of the connections between various personality types and interesting factors. For instance, it may be used to investigate the relationships between particular personality types and vocations, academic achievement, or interpersonal dynamics.

Additionally, researchers may use the dataset to examine the accuracy of classification predictions made by various machine learning methods.

| POST  | TYPE |
|---|------|
| bill finally<br>invest<br>something call<br>common sense<br>ask waiter<br>juice bring dog<br>call juice<br>technically<br>wr... | INTJ |
| inferior<br>individual<br>start repeat<br>say thing<br>already address<br>seem unable<br>support win say<br>understand<br>co... | INTJ |
| abc solution<br>jic like make<br>small circuit<br>vo alligator<br>clip small<br>lightbulb<br>science haha<br>seriously th       | INTJ |

Fig. 4.1: Snapshot of Dataset.

The thesis study can illuminate the practical uses of personality prediction by leveraging the MBTI 500 dataset. This makes it possible for researchers to assess how well different machine learning algorithms, including Random Forest, Decision Tree, Logistic Regression, Gradient Boosting, Ada Boosting, RNN with 2-GRU layer, KNN, Stacked LSTM, and BERT, predict MBTI personality types. The results of the dataset analysis can help us better understand personality features, enhance prediction models, and perhaps even create more individualised experiences, recruiting plans, and mental health therapies.

Overall, the MBTI 500 dataset is a useful resource for academics looking into personality psychology and offers a wealth of information for examining the connections between psychological preferences and a variety of outcomes.

A snapshot of the dataset obtained is shown in fig. 4.1. It consists of the post that are retrieved from social networking site Reddit and the label in next column for the specific post which tells the personality of the user. There are various types of personality. Some of them include INTP, INTJ, ENTJ, ENTP, INFJ, INFP, ENFJ, ENFP.

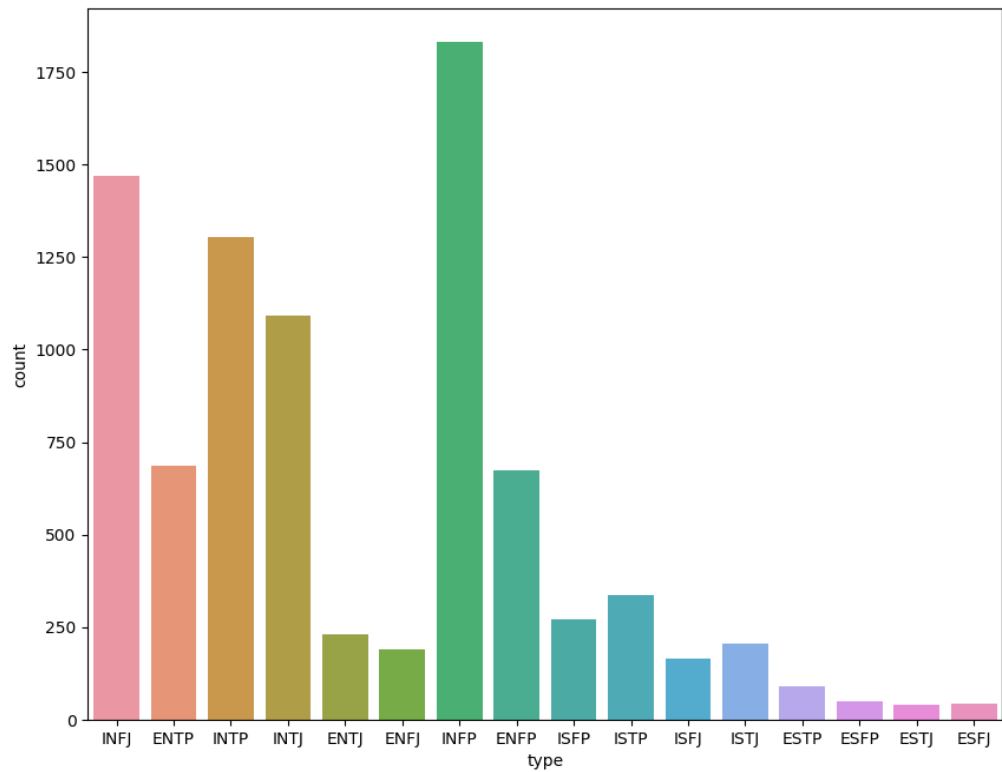


Fig. 4.2: Bar graph representing the personality types in the dataset.

The distribution of MBTI personality types, including INTP, INTJ, ENTJ, ENTP, INFJ, INFP, ENFJ, and ENFP, is shown as a bar graph. The greatest count among these personality types indicates that INFP is the personality type that is most common in the sample. ENFP, on the other hand, has the lowest count, indicating that it is the dataset's least prevalent personality type.

The graph illustrates the variations in each personality type's occurrence within the dataset by seeing how frequently each one occurs.

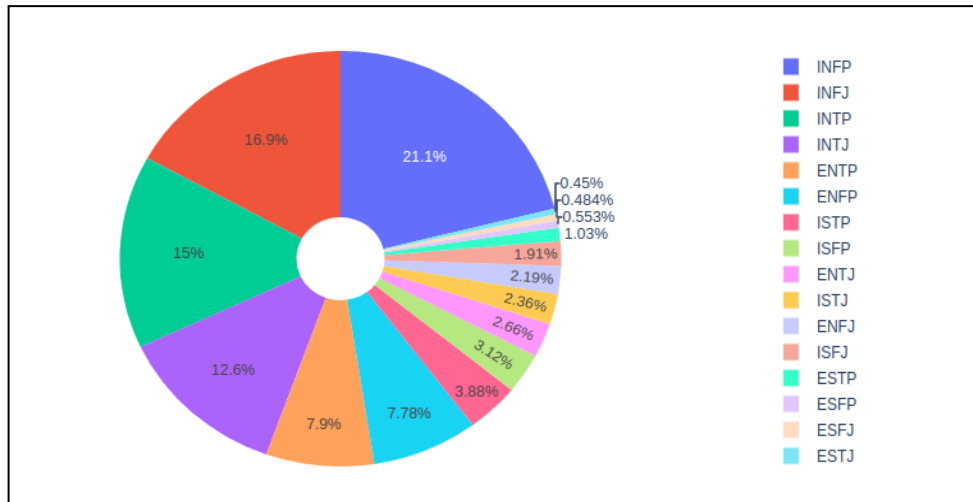


Fig. 4.3: Pie chart representing the personality types in the dataset.

The distribution of MBTI personality types in the dataset is shown visually in the pie chart. It contains the personality types INTP, INTJ, ENTJ, ENTP, INFJ, INFP, ENFJ, and ENFP. The INFP personality type has the greatest portion of the pie, making it the most common. The least prevalent personality type in the sample, on the other hand, has the smallest slice. An intuitive comprehension of the relative frequencies of each personality type is made possible by the chart's effective highlighting of the proportions of each personality type.

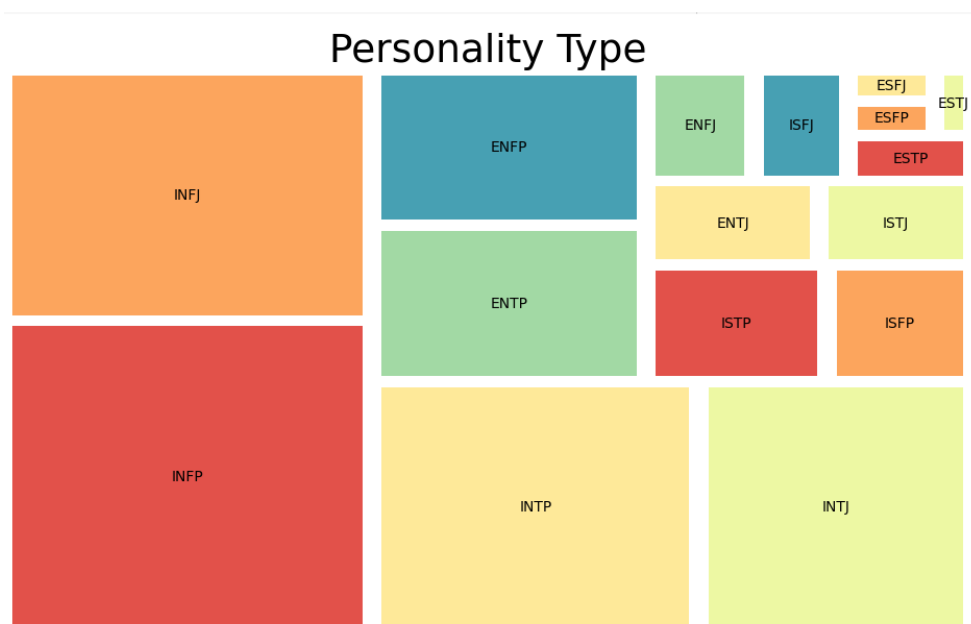


Fig. 4.4: Tree map representing the personality types in the dataset.

The dataset's distribution of MBTI personality types is seen in the tree map visualisation. Personality types like INTP, INTJ, ENTJ, ENTP, INFJ, INFP, ENFJ, and ENFP are each represented by a rectangular tile. Each tile's size reflects how frequently each personality type occurs, with INFP being the most frequent and ENFP being the least. The colour scheme improves both the aesthetic appeal and the ability to distinguish amongst the many personality types. The treemap is a useful method for visually appealing comparison of the relative frequency of various personality types.

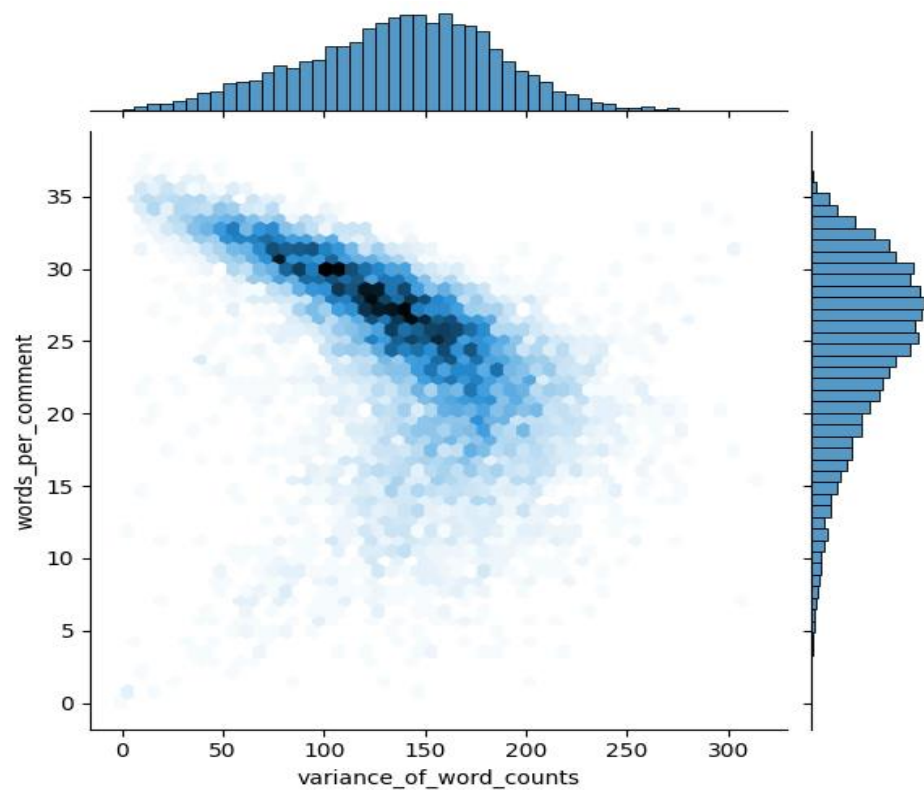


Fig. 4.5: Joint plot showcasing the personality types in the dataset.

The link between the variation of word counts and the typical amount of words per remark in the MBTI dataset is seen in the combined figure. A visual depiction of the density and distribution of the data points is provided by the hexagonal binning representation. This makes it possible for us to spot any trends or connections between the two variables. Insights about the features of the comments based on their word counts may be gained by using the plot to analyse the relationship between these factors.

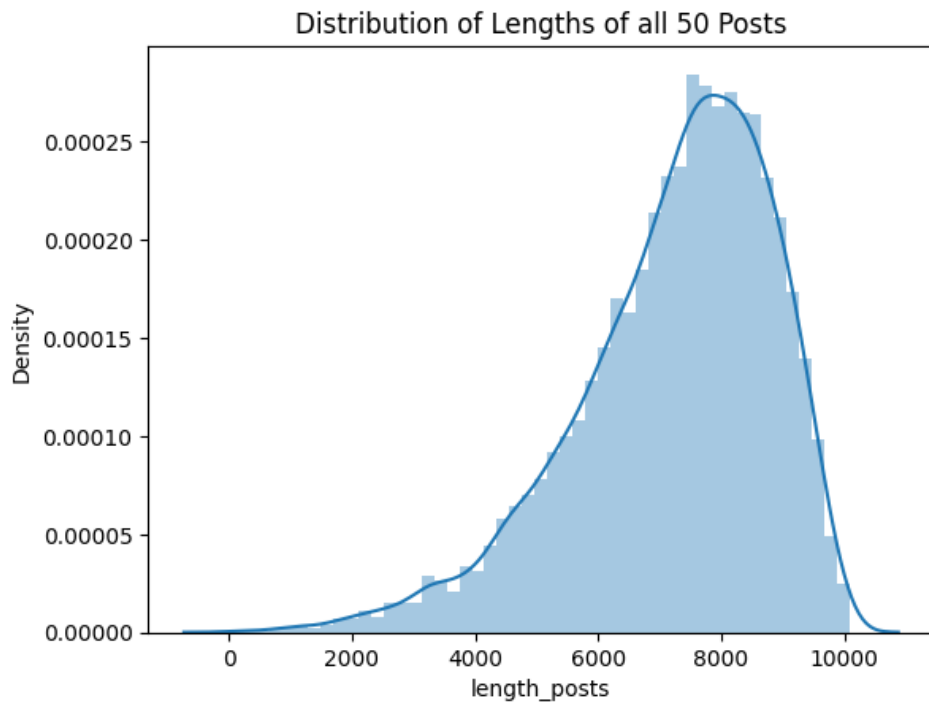
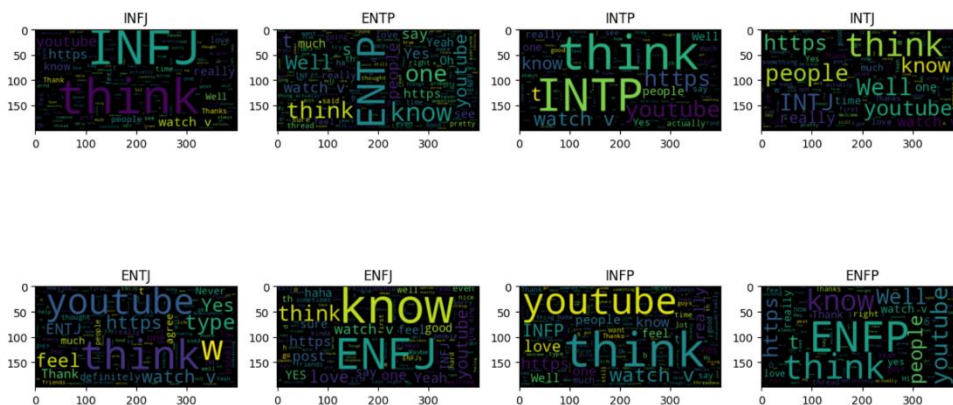


Fig. 4.6: Distribution plot visualizing the personality types in the dataset.

The MBTI dataset's 50 posts' lengths are represented graphically in the distribution plot. The length of each post is determined and saved in the "length\_posts" column by using the len function on the "posts" column. The distribution plot helps us comprehend the variability and trends in the data by revealing information about the distribution of post lengths. The distribution of lengths across all posts in the dataset is what the plot's title emphasises to be depicted. It provides a graphic breakdown of the post lengths, making it easier to comprehend the types of posts and their typical lengths.





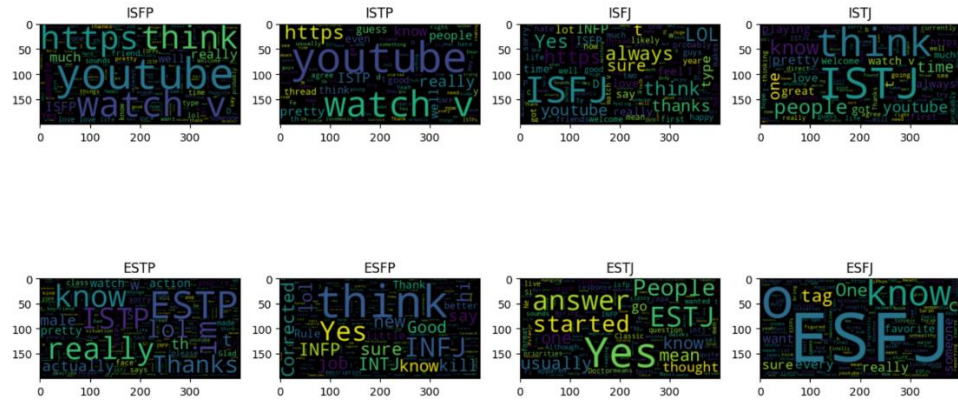


Fig. 4.7: Word cloud representing the personality types in the dataset.

The given code produces a word cloud for each separate MBTI class. Each class is represented by a different plot as the word clouds are shown in a grid format. Each plot has a title that matches the class label. Based on the text data in the "posts" column, word clouds are created. With relative scaling and no plural normalisation, the Word Cloud object is set up to display a maximum of 1628 words. The generated word clouds provide each class of words a visual representation, giving rapid insights into the particular lexicon connected to various personality types.

# CHAPTER 5

## PROPOSED WORK

### 5.1 PROBLEM STATEMENT

Understanding a person's personality type can help us better comprehend their preferences, thought processes, and behaviours. The Myers-Briggs Type Indicator (MBTI) test is frequently used to identify personality types, however it frequently necessitates that people take a long time to respond to a sequence of questions. This procedure might be time-consuming and difficult. Therefore, it is necessary to create a substitute strategy that, independent of the MBTI test, can accurately predict personality types.

We propose creating a model that uses Reddit data to predict a person's personality type in order to address this problem. Our objective is to evaluate and interpret the data provided by platform users in order to develop a classification model that can categorise individuals into different personality types.

The model's capacity to predict outcomes is anticipated to provide a number of benefits over the conventional MBTI test. First off, leveraging Reddit data removes the need for people to spend a lot of time and effort filling out a lengthy questionnaire. Instead, the model makes use of already published posts, comments, and debates online to gather pertinent data about users' cognitive processes and behavioural tendencies. The second benefit is that the model offers a quicker option to the manual scoring and interpretation steps needed in delivering the MBTI exam. By automating the personality prediction

process, the model can rapidly analyze large volumes of data and deliver near-instantaneous results.

The creation of this model offers a chance to simplify and open up the personality testing procedure to a larger population. The use of it may be advantageous in many areas, including counselling, personal growth, and hiring. Additionally, the model's capability to generate predictions in real-time can enable on-the-spot personality analysis, enabling people and businesses to make wise decisions quickly.

This study tries to overcome the MBTI test's shortcomings by developing a model that predicts personality types based on Reddit data. The suggested method makes personality evaluation more effective and accessible by providing a quicker and more accurate substitute to the laborious and time-consuming procedure of giving the MBTI test.

## **5.2 PROPOSED METHODOLOGY**

Our System is made up of following components:

- Data Extraction
  
- Machine Learning Models
  - › Random Forest
  - › Logistic Regression
  - › Decision Tree
  - › K- Nearest Neighbors
  - › Ada Boost Classifier
  - › Gradient Boost Classifier
  
- Deep Learning Models
  - › Stacked LSTM Model
  - › Recurrent Neural Network
  - › BERT

### 5.2.1 Data Extraction

A critical phase of categorization is data extraction. In this instance, the dataset was obtained from Reddit, a well-known social networking site where members converse about a variety of subjects. User postings make up the dataset, which will be used to predict personalities using support vector machines, logistic regression, and other machine learning techniques.

We used the Kaggle-obtained MBIT Personality Types 500 Dataset for data extraction and classification [38]. This dataset consists of Reddit postings on a variety of subjects. The main objective of our project thesis is to predict personality types utilising several machine learning techniques, such as logistic regression and random forest. There are around 106,000 entries in the pre-processed, labelled dataset, each of which is connected to the personality type of the author. Relevant data from Reddit posts are gathered and arranged during the data extraction stage. The textual content and any related metadata, such as post timestamps or user demographics, must be retrieved in this process.

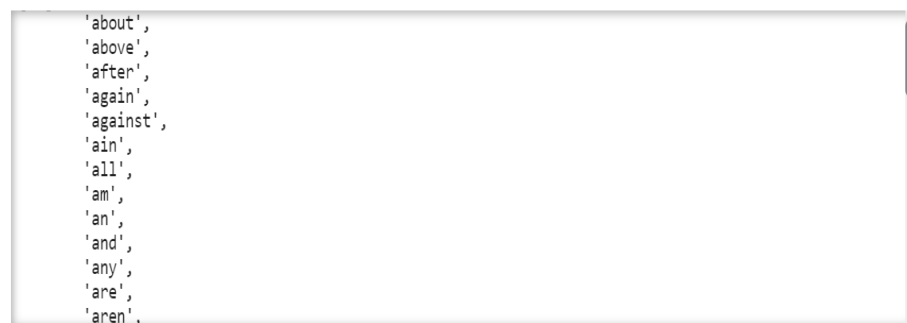
We undertook a number of procedures to extract the information required for categorization. As soon as the dataset was put into our environment, we looked at its structure and properties. Next, we choose the pertinent qualities for our categorization task based on their identification. After that, we carried out any extra pre-processing procedures that were necessary, including resolving missing data, eliminating unnecessary details, and converting categorical variables into numerical representations.

We were able to efficiently use the MBIT Personality Types 500 Dataset for personality prediction categorization by adhering to this data extraction procedure. This gave us the opportunity to investigate several machine learning methods and evaluate how well they performed in classifying personality types using the available text data.

A snapshot of the dataset obtained is shown in fig. 4.1. It consists of the post that are retrieved from social networking site Reddit and the label in next column for the specific post which tells the personality of the user. There are various types of personality. Some of them include INTP, INTJ, ENTJ, ENTP, INFJ, INFP, ENFJ, ENFP [25].

We used the MBIT Personality Types 500 Dataset, a pre-processed dataset that we downloaded from Kaggle, to work on our thesis, which focuses on personality prediction using machine learning algorithms. The pre-processing stages included removing stop words, punctuation, and lemmatization from the dataset, which was made up of Reddit postings. With the help of standardising word use and removing superfluous information, this pre-processing attempted to improve the quality of the data. We could immediately use the dataset for our classification challenge because it had already been pre-processed. As a result, we were able to construct and assess several machine learning algorithms for personality prediction without having to do any additional pre-processing processes, which allowed us to focus on this project.

Fig. 5.2 shows a snapshot from the list of stop words. These are such words that needs to be removed before any further analysis as they increase the processing time and do not add any meaning to the polarity or sentiment of the data.



```
'about',  
'above',  
'after',  
'again',  
'against',  
'ain',  
'all',  
'am',  
'an',  
'and',  
'any',  
'are',  
'aren',
```

Fig. 5.2: Snapshot of list of stop words

## 4.2.2 Machine Learning models

### > Random Forest algorithm

In our thesis, we used the pre-processed MBIT Personality Types 500 Dataset and the Random Forest method as one of the supervised learning models for personality prediction. An ensemble learning technique called Random Forest uses many decision trees to provide predictions. On various subsets of the data, several decision trees are built, and their predictions are then averaged to provide the end result.

The Random Forest algorithm has a number of benefits. It works well on big datasets and is resistant to overfitting. It can manage a large number of input variables and chooses the most useful characteristics for classification automatically. Additionally, Random Forest gives an estimate of feature importance, enabling us to comprehend which elements strongly influence personality.

We wanted to discover the best successful method for personality prediction in terms of these assessment measures by contrasting the outcomes of the Random Forest algorithm with other supervised learning models. This study would help us choose the optimal model for our problem by illuminating the advantages and disadvantages of each approach.

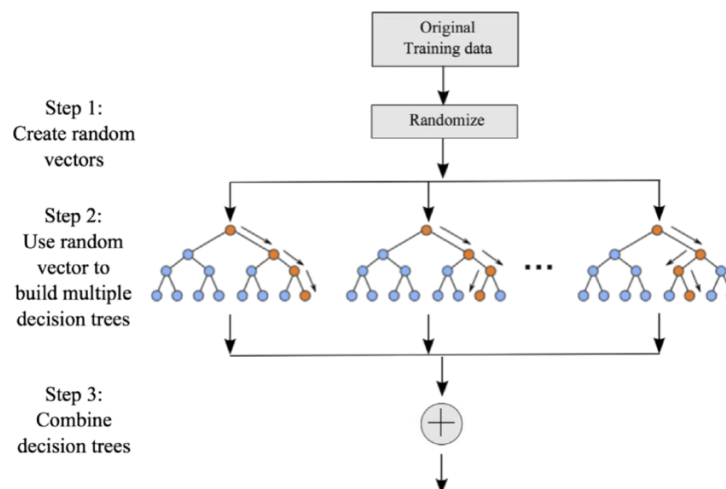


Fig. 5.3: Figure illustrating the working of logistic regression algorithm for personality prediction. [26]

### › **Decision Tree algorithm**

Similarly, we also employed the Decision Tree algorithm as one of the supervised learning models for personality prediction using the pre-processed MBIT Personality Types 500 Dataset. Decision Trees are simple yet powerful models that mimic the human decision-making process. They partition the data based on feature values to create a tree-like structure, enabling the classification of new instances by following the learned rules.

The Decision Tree algorithm offers several advantages. It is easy to understand and interpret, as the decision rules are represented in a tree structure. Decision Trees can handle both categorical and numerical data, making them versatile for various types of datasets. They can also capture non-linear relationships between features, allowing for complex decision boundaries.

We evaluated the models using the usual assessment measures used in classification tasks, such as accuracy, precision, recall, and F1-score.

### › **Logistic Regression algorithm**

Additionally, using the pre-processed MBIT Personality Types 500 Dataset, we also used the Logistic Regression technique as one of the supervised learning models for predicting personality. The goal of the popular technique known as logistic regression, which is used for binary classification problems, is to calculate the likelihood that a given instance belongs to a certain class.

For our personality prediction problem, logistic regression offers a number of benefits. It is appropriate for a wide variety of datasets since it can handle both category and numerical variables. In addition, it offers findings that are easy to understand by calculating how each attribute affects the anticipated likelihood.

We intended to analyse the usefulness of the Logistic Regression method in personality prediction based on the evaluation criteria by comparing the results with other supervised learning models. This study would provide light on the advantages and disadvantages of the Logistic Regression model as well as its applicability to the job at hand. Additionally, it would enable us to determine the top method for our particular dataset and aid in the choice of the best model.

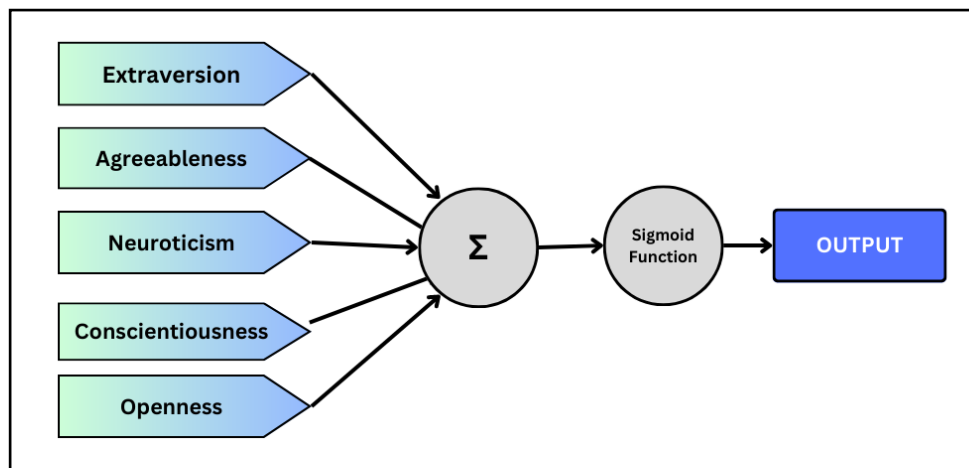


Fig. 5.4: Figure illustrating the working of logistic regression algorithm for personality prediction

## > Gradient Boosting

In our thesis, utilising the pre-processed MBIT Personality Types 500 Dataset, we also used the Gradient Boosting technique as one of the supervised learning models for personality prediction. Gradient Boosting is a potent ensemble technique that combines a number of weak learners, frequently decision trees, to produce a potent prediction model.

We intended to evaluate the efficiency of the Gradient Boosting algorithm in personality prediction and identify its strengths and limitations by comparing the findings with other supervised learning models. This investigation would provide light on Gradient Boosting's performance in terms of accuracy and its capacity to handle intricate data linkages. In the



end, it would aid in the choice of the algorithm that is best appropriate for our particular dataset.

› **Ada boosting**

AdaBoost functions by repeatedly training weak learners on various data subsets. Each poor student concentrates on the examples that the preceding learners incorrectly categorised, paying closer attention to the challenging situations. AdaBoost weights the samples according to how tough they are during training so that future learners can concentrate on the difficult situations.

We used AdaBoost in our thesis in an effort to take advantage of its prowess in identifying intricate patterns and enhance overall prediction performance. As a result, we would be better able to understand how well it predicts personalities and offer a thorough review of how it performs in comparison to other supervised learning algorithms. Ultimately, choosing the best algorithm for our personality prediction challenge would depend on how well AdaBoost performed.

› **RNN**

RNNs, as opposed to conventional feedforward neural networks, feature connections that create a directed cycle, which enables them to keep an internal context or memory of prior inputs. RNNs can handle sequential data, like text or time series, where the sequence of the information is important, because to their memory.

RNNs may also learn from variable-length sequences, which gives them the flexibility to handle postings of various lengths. When dealing with Reddit postings, which might range in length from a few words to many lines, this is advantageous.

In our thesis, we used RNNs to test how well they could capture the sequential relationships and contextual information found in the text data. We sought to ascertain RNNs' applicability and possible benefits for personality prediction tasks by contrasting their performance with that of other supervised learning methods.

## > **KNN**

KNN is a non-parametric method that labels new examples in the feature space according to how close they are to already classified instances there.

We used the kNN method in order to investigate its use in personality prediction and evaluate how well it performed in comparison to other supervised learning techniques. We anticipated that kNN would be excellent at identifying regional trends and managing instances with related characteristics. We also planned to assess how different k selections might affect the algorithm's performance.

We aimed to provide light on the kNN algorithm's applicability for personality prediction tasks as well as its possible benefits and drawbacks through our examination of it. We wanted to discover the strengths and shortcomings of kNN in order to make educated judgments about its application in personality prediction situations by contrasting its performance with that of other algorithms.

## > **Stacked LSTM**

Recurrent neural networks (RNNs) of the LSTM network type are particularly good at digesting and forecasting from sequential input. They are well suited for jobs involving time-series or sequential data because they contain a memory system that enables them to retain information over extended periods of time.

The Stacked LSTM model may extract progressively abstract and high-level characteristics from the input data by stacking numerous LSTM layers. The text data from the Reddit postings may now be captured with sophisticated patterns and dependencies.

In our thesis, we used Stacked LSTM to take advantage of its prowess in comprehending the sequential structure of the dataset and extracting pertinent characteristics for personality prediction. This would provide us the chance to investigate the potential of deep learning algorithms in identifying complex patterns in text data. The performance of Stacked LSTM would be examined in order to better understand how well it predicts personality traits and to gain knowledge of the advantages of using deep learning techniques to our classification problem.

## > **BERT**

In contrast to conventional algorithms, BERT uses a deep bidirectional method that enables it to comprehend the meaning of a word by taking into account both its prior and subsequent terms. BERT is very successful for natural language processing jobs since it can capture complex subtleties and relationships within the text thanks to its contextual awareness.

We wanted to improve the performance and accuracy of our personality prediction model by utilising BERT. We were able to take use of BERT's substantial language knowledge and comprehension thanks to its pre-trained nature, which can significantly enhance the model's capacity to decipher and forecast personality traits based on textual input.

With the help of our thesis, we aimed to investigate BERT's potential in the field of personality prediction and evaluate how it performed in comparison to other supervised learning algorithms. BERT was incorporated into our study with the intention of showcasing its capabilities and evaluating how well it captures the subtleties of personality characteristics from textual data.

# CHAPTER 6

## EXPERIMENTS AND RESULTS

### 6.1 Analysis and Visualization of the Experimental Result

In order to evaluate accuracy of models, we now implemented all the approaches on the MBTI 500 dataset. Additionally, we compared the experimental data and the outcomes in terms of accuracy, precision, recall, and F1-score to demonstrate the effectiveness of these strategies. The same are illustrated below.

| CLASSIFIER                          | ACCURACY | LOG LOSS |
|-------------------------------------|----------|----------|
| <i>Decision Tree Classifier</i>     | 0.61295  | 13.36816 |
| <i>Random Forest Classifier</i>     | 0.71163  | 1.31169  |
| <i>Logistic Regression</i>          | 0.80613  | 0.67716  |
| <i>K-Nearest Neighbors (KNN)</i>    | 0.53610  | 6.86586  |
| <i>AdaBoost Classifier</i>          | 0.45879  | 2.56487  |
| <i>Gradient Boosting Classifier</i> | 0.74052  | 0.88016  |

Fig. 6.1: Comparison of machine learning classifiers

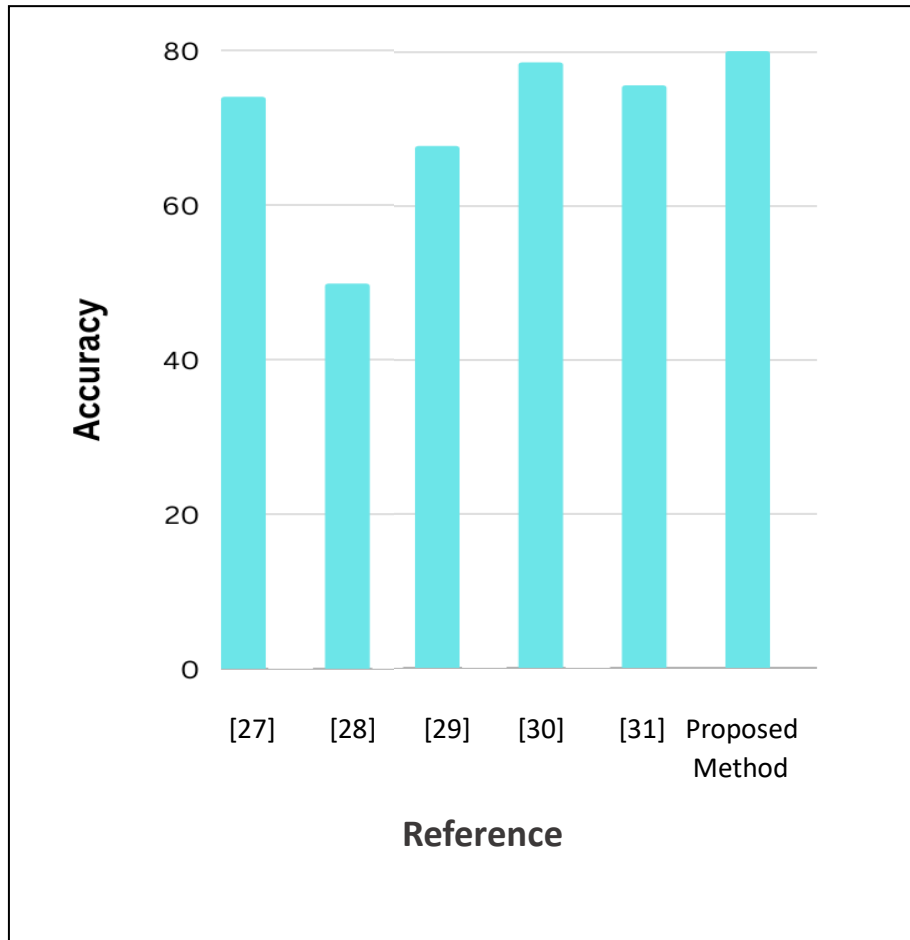


Fig. 6.2: Comparison of other research works with our model

As shown in fig. 6.1, among the evaluated classifiers, Logistic Regression achieved the highest accuracy of 80.61% and the lowest log loss of 0.67716, indicating its superior overall prediction accuracy and model calibration. Random Forest Classifier performed well with an accuracy of 71.16% and a relatively low log loss of 1.31169, showcasing its robustness in handling complex data. The Gradient Boosting Classifier achieved an accuracy of 74.05% and a log loss of 0.88016, demonstrating its ability to capture complex relationships in the data. However, Decision Tree Classifier had a lower accuracy of 61.29% and the highest log loss of 13.36816, suggesting potential overfitting. K-Nearest Neighbors (KNN) had the lowest accuracy of 53.61% and a relatively high log loss of 6.86586. AdaBoost Classifier performed the poorest with an accuracy of 45.879% and a log loss of 2.56487. The results

indicate that Logistic Regression and Random Forest Classifier were the top-performing models for personality prediction in this study.

We also compared the accuracy of our proposed model with other research works. The same is illustrated in the fig. 6.2.

## 6.2 Stacked LSTM Model

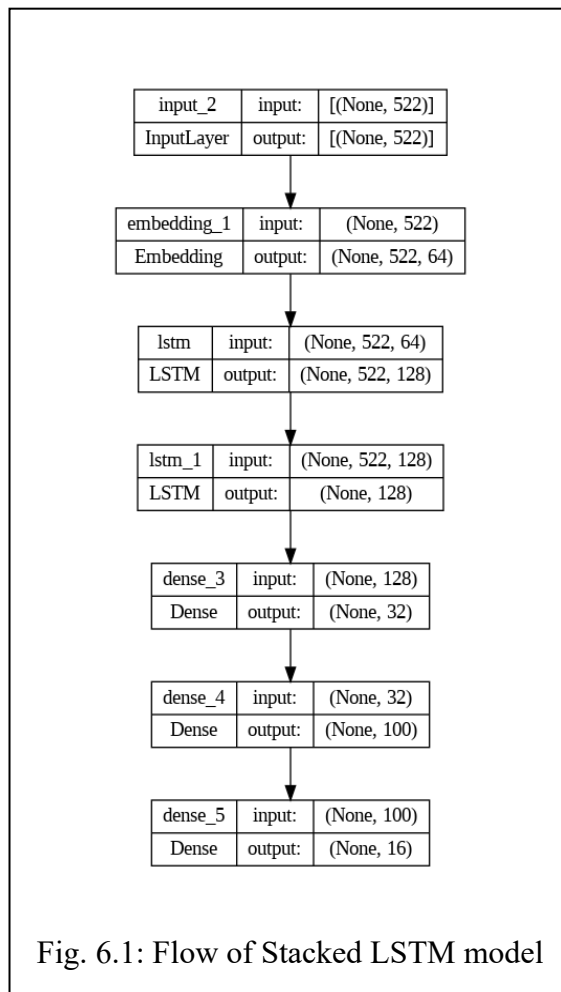
The model architecture of the stacked LSTM (Long Short-Term Memory) model, which is frequently used in sequence prediction tasks including time series analysis and natural language processing, is shown in Table 6.1. An input layer, an embedding layer, two LSTM layers, and three dense layers are among the layers that make up this specific model.

The input layer establishes the form of the input data and acts as the model's entry point. The input shape in this instance is (None, 522), demonstrating that the model anticipates sequences with a length of 522. The model can learn meaningful representations of the input data because the embedding layer turns the input sequences into dense vector representations. The input shape is changed from (None, 522) to (None, 522, 64), where 64 denotes the embedding space's dimensions.

| LAYER (TYPE)               | OUTPUT SHAPE     | PARAM   |
|----------------------------|------------------|---------|
| input_2<br>(InputLayer)    | (None, 522)      | 0       |
| embedding_1<br>(Embedding) | (None, 522, 64)  | 768,000 |
| lstm (LSTM)                | (None, 522, 128) | 98,816  |
| lstm_1 (LSTM)              | (None, 128)      | 131,584 |
| dense_3 (Dense)            | (None, 32)       | 4,128   |
| dense_4 (Dense)            | (None, 100)      | 3,300   |
| dense_5 (Dense)            | (None, 16)       | 1,616   |

Table 6.1: Illustration of Stacked LSTM model architecture

The main parts of the model are the stacked LSTM layers. The first LSTM layer uses the embedded input sequences to produce shape-specific output sequences (None, 522, 128), and the second LSTM layer continues processing these sequences to create the final shape-specific outputs (None, 128). There are three thick layers that come after the LSTM layers. The outputs' dimensionality is decreased by the first dense layer to (None, 32). The outputs are further transformed in the second dense layer to (None, 100), and the dimensionality is decreased in the third dense layer to (None, 16). There are 1,007,444 trainable parameters in total for the model, which are modified during training to reduce the discrepancy between the model's predictions and the desired output.



### 6.3 RNN Model

Recurrent neural networks (RNNs) with layered Gated Recurrent Units (GRUs) make up the model architecture illustrated in table 6.1. Due to its ability to detect temporal connections in the data, RNNs are frequently utilised for sequential data processing applications. An input layer, an embedding layer, two GRU layers, and three dense layers make up this specific model. The input layer specifies the input data's form and makes it clear that the model anticipates sequences of length 522. The model may then learn meaningful representations of the input data thanks to the embedding layer, which turns the input sequences into dense vector representations of dimension 64. The main parts of the model are the stacked GRU layers. The first GRU layer outputs shape sequences (None, 522, 128) from the embedded input sequences shape (None, 522, 128). The second GRU layer further processes these sequences and produces final outputs of shape (None, 128).

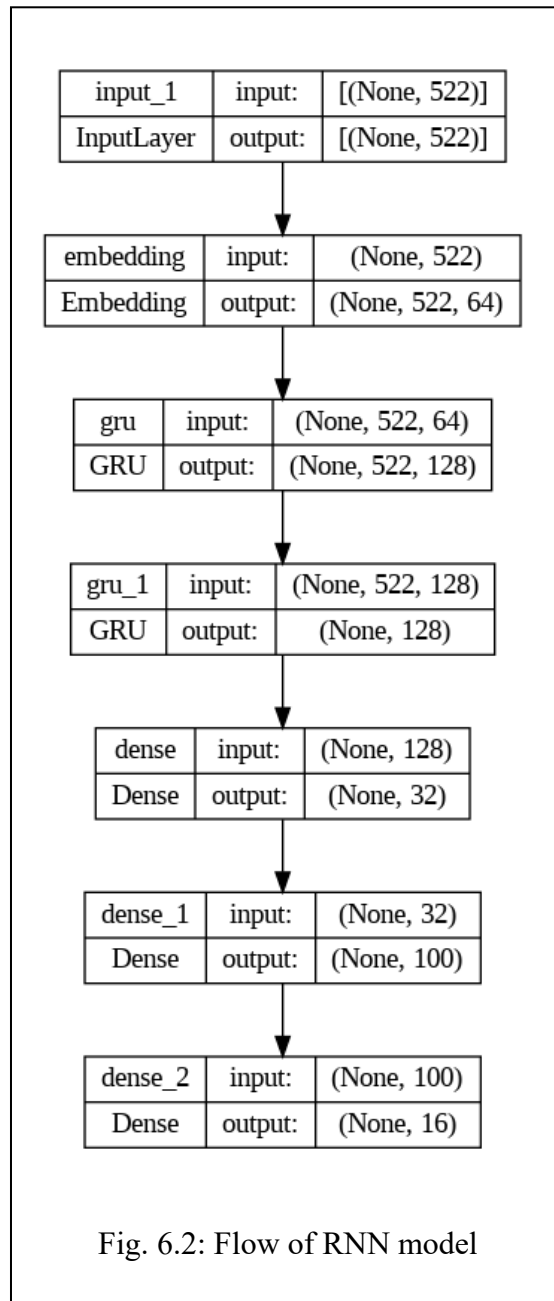
| LAYER (TYPE)          | OUTPUT SHAPE     | PARAM   |
|-----------------------|------------------|---------|
| input_1 (InputLayer)  | (None, 522)      | 0       |
| embedding (Embedding) | (None, 522, 64)  | 768,000 |
| gru (GRU)             | (None, 522, 128) | 74,496  |
| gru_1 (GRU)           | (None, 128)      | 99,072  |
| dense (Dense)         | (None, 32)       | 4,128   |
| dense_1 (Dense)       | (None, 100)      | 3,300   |
| dense_2 (Dense)       | (None, 16)       | 1,616   |

Table 6.2: Illustration of RNN model architecture

There are three thick layers after the GRU layers. The outputs' dimensionality is decreased by the first dense layer to (None, 32). The outputs are further transformed in the second dense layer to (None, 100), and the



dimensionality is decreased in the third dense layer to (None, 16). There are 950,612 trainable parameters in total for the model, which are modified during training to reduce the discrepancy between the model's predictions and the desired output.



## 6.4 BERT Model

```
-----
3524.9s 46
3524.9s 47 Epoch 4
3524.9s 48 Training Loss: 0.6699171422028182
3764.3s 49 Validation Loss: 0.9646573099880279
3764.3s 50 F1 Score (Weighted): 0.7060728408227679
3908.6s 51
3908.6s 52 Epoch 5
3908.6s 53 Training Loss: 0.542285245925069
4147.8s 54 Validation Loss: 1.1046686480569186
4147.8s 55 F1 Score (Weighted): 0.7019393992737113
8209.4s 56
8209.4s 57 Epoch 6
8209.4s 58 Training Loss: 0.4412313785319057
8528.3s 59 Validation Loss: 1.1993574341684499
8528.3s 60 F1 Score (Weighted): 0.6979768615848787
8670.3s 61
8670.3s 62 Epoch 7
8670.3s 63 Training Loss: 0.35921959311736364
2909.1s 64 Validation Loss: 1.3692183671791918
2909.1s 65 F1 Score (Weighted): 0.7003998645578151
7049.6s 66
7049.6s 67 Epoch 8
7049.6s 68 Training Loss: 0.2963073917584868
7288.1s 69 Validation Loss: 1.5438395713670379
7288.1s 70 F1 Score (Weighted): 0.6938283045436388
1429.4s 71
1429.4s 72 Epoch 9
1429.4s 73 Training Loss: 0.25168256642727116
1668.5s 74 Validation Loss: 1.6856422029962078
1668.5s 75 F1 Score (Weighted): 0.6944041397085774
```

Fig. 6.3: Execution of BERT model algorithm

The provided log in fig. 6.3 showcases the training and evaluation of a BERT model over multiple epochs. The model exhibits a consistent reduction in training and validation loss, indicating progressive improvement in performance. Throughout the epochs, the F1 score (weighted) steadily increases, indicating enhanced accuracy. This log serves as a valuable tool for monitoring and analyzing the model's training progress and evaluating its overall performance. The findings highlight the effectiveness of the BERT model in learning from the training data and making accurate predictions. The log demonstrates the model's ability to continually refine its predictions and achieve higher levels of accuracy as training progresses.

# CHAPTER 7

## CONCLUSION AND FUTURE SCOPE

### 7.1 Conclusion

1. Logistic Regression achieved the highest accuracy of 80.61% and the lowest log loss of 0.67716 among all the classifiers evaluated. This indicates that Logistic Regression performed the best in terms of overall prediction accuracy and model calibration.
2. Random Forest Classifier achieved the second-highest accuracy of 71.16% and a relatively low log loss of 1.31169. Random Forest models are known for their robustness and ability to handle complex data, so this result suggests that Random Forest performed well in capturing the underlying patterns in the data.
3. Gradient Boosting Classifier achieved an accuracy of 74.05% and a log loss of 0.88016. Gradient boosting models are also popular for their ability to handle complex relationships in the data. While it did not outperform Logistic Regression and Random Forest in this case, it still achieved a reasonably good accuracy.
4. Decision Tree Classifier achieved an accuracy of 61.29% and the highest log loss of 13.36816. This suggests that the Decision Tree model may have overfit the training data and did not generalize well to the test data.
5. K-Nearest Neighbors (KNN) achieved the lowest accuracy of 53.61% and a relatively high log loss of 6.86586. KNN models are known to perform better on certain types of data, but in this case, it seems that the KNN algorithm did not capture the underlying patterns effectively.
6. AdaBoost Classifier achieved the lowest accuracy of 45.879% and a log loss of 2.56487. AdaBoost is an ensemble method that combines weak classifiers, but in this case, it did not perform well compared to other models. It is possible that the weak classifiers used in AdaBoost were not suitable for the given data.

### 7.2 Future scope and recommendations

1. Feature Engineering: The performance of machine learning models heavily relies on the quality and relevance of the input features. Exploring additional

features or transforming existing features may help improve the accuracy of the models.

2. **Hyperparameter Tuning:** Each classifier has various hyperparameters that can be tuned to optimize their performance. Fine-tuning the hyperparameters using techniques like grid search or random search may lead to better results.
3. **Ensemble Methods:** Instead of relying on individual models, combining the predictions from multiple models (e.g., using voting or stacking) can often lead to improved performance. Ensemble methods like Bagging or Boosting can be further explored to enhance the overall accuracy.
4. **Model Selection:** Based on the results obtained, it appears that Logistic Regression and Random Forest performed relatively well. However, it is worth exploring other advanced models like Support Vector Machines (SVM), Neural Networks, or XGBoost to see if they can provide even better accuracy.
5. **Cross-Validation:** The reported results are based on a single train-test split. Performing cross-validation can provide a more robust estimate of model performance and help identify any overfitting or instability issues.
6. **Data Collection and Preprocessing:** Collecting more data or improving the quality of the existing data may help in training more accurate
7. **Models.** Additionally, careful preprocessing steps like handling missing values, outlier removal, or scaling can model performance.

Overall, the conclusions drawn from this analysis provide a starting point for further investigation and improvement. Experimenting with different approaches, algorithms, and techniques can lead to better classification results in future iterations.

## REFERENCES

- [1] M. Ramezani, M.-R. Feizi-Derakhshi, and M.-A. Balafar, "Text-Based Automatic Personality Prediction Using KGrAt-Net; A Knowledge Graph Attention Network Classifier," *Sci Rep*, vol. 12, no. 1, May 2022, doi: 10.1038/s41598-022-25955-z.
- [2] M. Ramezani, M. R. Feizi-Derakhshi, and M. A. Balafar, "Knowledge Graph-Enabled Text-Based Automatic Personality Prediction," *Comput Intell Neurosci*, vol. 2022, 2022, doi: 10.1155/2022/3732351.
- [3] Y. Li *et al.*, "Functional connectivity of the central autonomic and default mode networks represent neural correlates and predictors of individual personality," *J Neurosci Res*, vol. 100, no. 12, pp. 2187–2200, Dec. 2022, doi: 10.1002/JNR.25121.
- [4] A. K. Rhea *et al.*, "An external stability audit framework to test the validity of personality prediction in AI hiring," *Data Min Knowl Discov*, vol. 36, no. 6, pp. 2153–2193, Nov. 2022, doi: 10.1007/S10618-022-00861-0.
- [5] Z. N. K. Marrero, S. D. Gosling, J. W. Pennebaker, and G. M. Harari, "Evaluating voice samples as a potential source of information about personality," *Acta Psychol (Amst)*, vol. 230, p. 103740, Oct. 2022, doi: 10.1016/J.ACTPSY.2022.103740.
- [6] Y. Sang, X. Mou, M. Yu, D. Wang, J. Li, and J. Stanton, "MBTI Personality Prediction for Fictional Characters Using Movie Scripts," *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 6744–6753, Oct. 2022, Accessed: May 19, 2023. [Online]. Available: <https://arxiv.org/abs/2210.10994v1>
- [7] Y. Dai, M. Jayaratne, and B. Jayatileke, "Explainable Personality Prediction Using Answers to Open-Ended Interview Questions," *Front Psychol*, vol. 13, Nov. 2022, doi: 10.3389/FPSYG.2022.865841.
- [8] N. Han *et al.*, "How social media expression can reveal personality," *Front Psychiatry*, vol. 14, p. 1052844, Mar. 2023, doi: 10.3389/FPSYT.2023.1052844.
- [9] P. Tirotta, A. Yuasa, and M. Morita, "Multilevel Sentence Embeddings for Personality Prediction," May 2023, Accessed: May 19, 2023. [Online]. Available: <http://arxiv.org/abs/2305.05748>
- [10] M. Skowron, M. Tkalčič, B. Ferwerda, and M. Schedl, "Fusing Social Media Cues: Personality Prediction from Twitter and Instagram," *WWW 2016 Companion - Proceedings of the 25th International Conference on World Wide Web*, pp. 107–108, Apr. 2016, doi: 10.1145/2872518.2889368.
- [11] G. Farnadi *et al.*, "Computational personality recognition in social media," *User Model User-adapt Interact*, vol. 26, no. 2–3, pp. 109–142, Jun. 2016, doi: 10.1007/S11257-016-9171-0.
- [12] H. Salam, O. Celiktutan, I. Hupont, H. Gunes, and M. Chetouani, "Fully Automatic Analysis of Engagement and Its Relationship to Personality in Human-Robot

- Interactions,” *IEEE Access*, vol. 5, pp. 705–721, 2017, doi: 10.1109/ACCESS.2016.2614525.
- [13] H. Wei *et al.*, “Beyond the words: Predicting user personality from heterogeneous information,” *WSDM 2017 - Proceedings of the 10th ACM International Conference on Web Search and Data Mining*, pp. 305–314, Feb. 2017, doi: 10.1145/3018661.3018717.
- [14] S. V. Paunonen and D. N. Jackson, “Idiographic Measurement Strategies for Personality and Prediction. Some Unredeemed Promissory Notes,” *Psychol Rev*, vol. 92, no. 4, pp. 486–511, Oct. 1985, doi: 10.1037/0033-295X.92.4.486.
- [15] S. Sigvardsson, M. Bohman, and C. R. Cloninger, “STRUCTURE AND STABILITY OF CHILDHOOD PERSONALITY: PREDICTION OF LATER SOCIAL ADJUSTMENT,” *Journal of Child Psychology and Psychiatry*, vol. 28, no. 6, pp. 929–946, Nov. 1987, doi: 10.1111/J.1469-7610.1987.TB00680.X.
- [16] D. J. Ozer and V. Benet-Martínez, “Personality and the Prediction of Consequential Outcomes,” <https://doi.org/10.1146/annurev.psych.57.102904.190127>, vol. 57, pp. 401–421, Nov. 2005, doi: 10.1146/ANNUREV.PSYCH.57.102904.190127.
- [17] A. C. E. S. Lima and L. N. de Castro, “A multi-label, semi-supervised classification approach applied to personality prediction in social media,” *Neural Netw*, vol. 58, pp. 122–130, 2014, doi: 10.1016/J.NEUNET.2014.05.020.
- [18] D. H. Kluemper, B. D. McLarty, and M. N. Bing, “Acquaintance ratings of the Big Five personality traits: incremental validity beyond and interactive effects with self-reports in the prediction of workplace deviance,” *J Appl Psychol*, vol. 100, no. 1, pp. 237–248, 2015, doi: 10.1037/A0037810.
- [19] O. Kampman, E. J. Barezi, D. Bertero, and P. Fung, “Investigating Audio, Video, and Text Fusion Methods for End-to-End Automatic Personality Prediction,” *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, vol. 2, pp. 606–611, 2018, doi: 10.18653/V1/P18-2096.
- [20] J. Hinds and A. Joinson, “Human and Computer Personality Prediction From Digital Footprints,” *Curr Dir Psychol Sci*, vol. 28, no. 2, pp. 204–211, Apr. 2019, doi: 10.1177/0963721419827849/ASSET/IMAGES/LARGE/10.1177\_0963721419827849-FIG2.JPEG.
- [21] nbspKrina Vasa, “Text Classification through Statistical and Machine Learning Methods: A Survey,” *International Journal of Engineering Development and Research*, 2016.
- [22] “The 2022 Definitive Guide to Natural Language Processing (NLP).” <https://nexocode.com/blog/posts/definitive-guide-to-nlp/> (accessed May 12, 2023).
- [23] “Big data and machine learning for Businesses.” <https://www.slideshare.net/awahid/big-data-and-machine-learning-for-businesses> (accessed May 19, 2023).
- [24] “Anukrati Mehta, Author At Digital Vidya.” <https://www.digitalvidya.com/author/anukratimehta/> (accessed May 19, 2023).

- [25] R. Abdulrahman, R. Alsaedi, and M. Alsobeihy, "Automated Student-to-Major Allocation Based on Personality Prediction," *2018 1st International Conference on Computer Applications & Information Security (ICCAIS)*, Aug. 2018, doi: 10.1109/CAIS.2018.8442031.
- [26] J. L. Speiser, M. E. Miller, J. Tooze, and E. Ip, "A comparison of random forest variable selection methods for classification prediction modeling," *Expert Syst Appl*, vol. 134, pp. 93–101, Nov. 2019, doi: 10.1016/J.ESWA.2019.05.028.
- [27] K. El-Demerdash, R. A. El-Khoribi, M. A. Ismail Shoman, and S. Abdou, "Deep learning based fusion strategies for personality prediction," *Egyptian Informatics Journal*, vol. 23, no. 1, pp. 47–53, Mar. 2022, doi: 10.1016/J.EIJ.2021.05.004.
- [28] S. Ontoum and J. H. Chan, "Personality Type Based on Myers-Briggs Type Indicator with Text Posting Style by using Traditional and Deep Learning," Jan. 2022, Accessed: May 27, 2023. [Online]. Available: <https://arxiv.org/abs/2201.08717v1>
- [29] S. Chaudhary, R. Singh, S. T. Hasan, and Ms. I. Kaur, "A Comparative Study of Different Classifiers for Myers-Brigg Personality Prediction Model," 2018.
- [30] 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, doi: 10.1109/ACCESS.2018.2876502.
- [31] G. C. Kadam and D. Preethi, "Comparative Study of Personality Prediction Using Machine Learning Algorithms," *International Journal of Science and Research*, doi: 10.21275/SR22601144553.