# Personality Prediction And Group Detection Using Social Media Posts

Muskan Kuchhal
*Dept. of Computer Science & Engg.*
*Delhi Technological University*
Delhi, India
muskankuchhal_2k18co217@dtu.ac.in

Puneet Jangid
*Dept. of Computer Science & Engg.*
*Delhi Technological University*
Delhi, India
puneetjangid_2k18co263@dtu.ac.in

Muskan Saini
*Dept. of Computer Science & Engg.*
*Delhi Technological University*
Delhi, India
muskansaini_2k18co218@dtu.ac.in

Rajni Jindal
*Dept. of Computer Science & Engg.*
*Delhi Technological University*
Delhi, India
rajnijindal@dce.ac.in

2022 IEEE 7th International conference for Convergence in Technology (I2CT) | 978-1-6654-2168-3/22/$31.00 ©2022 IEEE | DOI: 10.1109/I2CT54291.2022.9825280

***Abstract*—Personality is an essential component of our every day life. It has an impact on how we live, speak, respond, and express ourselves, as well as on our mental health. Personality analysis is a natural human ability that is utilised every day with a wide range of individuals and for a wide range of objectives. Personality profiling, in particular, has a wide range of real-world applications, such as mental disease screening exams, human resource interview shortlisting, friend recommendations, and writer recommendations based on the interplay of personalities that people like reading about. Kaggle's (MBTI) Myers Briggs Personality Type Dataset will be used for the project. The Myers–Briggs Type Indicator (MBTI) is currently widely used and recognised as one of the most reliable instruments. In this work, a comparative study is performed for predicting MBTI personality types using different machine learning algorithm . The results obtained after a series of progressive techniques are compared on the basis of various metrics and the findings reveal that both in terms of accuracy and dependability the performance has been improved as compared to previous works the solution deployment has been detailed, and a large number of feature vector combinations and machine learning models were compared.**

***Index Terms*—Personality prediction , Machine learning, Myers–Briggs Type Indicator**

## I. Introduction

Personality, in general terms, refers to a person's distinctive patterns of thoughts, feelings, and behaviour. Although earlier this concept of importance of personality of people was considered trivial but in the present times, as population bombarded, personality types of people reveal a lot more use such as mental health and psychology of people, establishing friendships and social groups of similar interests and even guiding people to pursue careers that might seem to be most suitable to them.

Understanding personality types has applications in the job. It can assist with our leadership style, resolving conflicts more effectively, communicating more effectively, understanding how others make decisions, coaching others, improving sales abilities, and retaining important personnel.

## II. Related Work

In [3] a significant contribution has been made to the field of personality prediction and some of them have taken psychological tests into account for deciding the personalities of individuals while others have used machine learning algorithms like Naïve Bayes analyses troll traits and authorises users on various Twitter platforms Techniques employed include sequential mining optimization, random forest, and Naive Bayes algorithms. The dataset for this study is based on users who are and are not impacted by trolling. The study indicated that automated categorization is beneficial in the detecting procedure.

In [4] They used the MyPersonality dataset and employed linear regression and support vector regression to estimate the personality of Facebook users. According to the findings of this study, linear regression is a better alternative for prediction.

The goal of [7] is to locate marijuana-related Facebook posts uploaded by people from various backgrounds and assess sentiment based on their responses to the postings. Data is collected from the High Times Magazine Facebook page for this work, and features are recorded. NLP processing, like tokenization,removing stop words and stemming the text, is applied to the retrieved posts. Finally, negative and positive attitudes are established based on word use.

[6] has used Multinomial Nave Bayes, AdaBoost, and LDA to determine which method is most relevant. According to their findings, Multinomial Nave Bayes has the best accuracy. They predicted personality using a real-time Twitter dataset.

In [5] the use of several machine learning algorithms on the MBTI personality dataset to categorise the text into distinct personality types is the primary goal. KNN, Decision Tree, Random Forest, MLP, Logistic Regression, SVM, XGBoost, MNB, and Stochastic Gradient Descent techniques were utilised to get the findings, and a comparison research was conducted. According to the study, the XGboost approach produces the greatest outcomes.

[1] created a machine learning model utilising the natural language processing toolbox and the XGBoost classifier for training four binary classifiers for each of the personality type axis The training and validation data were split into 7:3 ratio. The model's accuracy was measured to assess performance on the testing set.

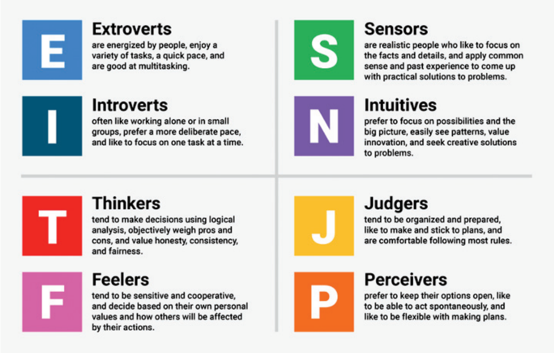
The research in [2] focuses on assessing the effectiveness of several classifiers on the MBTI dataset for personality prediction. The Naive Bayes, Logistic Regression, SVM, and Random Forests models are used, and parameter adjustment is performed to increase the accuracy of the Logistic Regression model. The classifiers' accuracy and F Measure were employed as scoring metrics.

## III. Background

This section provides the information about MBTI (Myers Briggs Personality Type Indicator) and personality types.

### A. Personality Types

Personality implies defining an individual's behaviour or character. Personality is defined as the distinctive sets of cognition, behaviours, and passionate examples that emerge from biological and ecological components. It reflects the people's differences in thoughts, behaviour, and emotions. Personality traits are constant in the universe because they provide a sense of high and low of explicit features in an individual on a consistent quality rather than exhibiting a specific personality. Personality differences should be expected to prevent the cause of deterrents or friction in the subject of work or education. In this manner, an individual should be distinguished in the enrolment of schools or working environment.



**Extroverts** are energized by people, enjoy a variety of tasks, a quick pace, and are good at multitasking.

**Introverts** often like working alone or in small groups, prefer a more deliberate pace, and like to focus on one task at a time.

**Thinkers** tend to make decisions using logical analysis, objectively weigh pros and cons, and value honesty, consistency, and fairness.

**Feelers** tend to be sensitive and cooperative, and decide based on their own personal values and how others will be affected by their actions.

**Sensors** are realistic people who like to focus on the facts and details, and apply common sense and past experience to come up with practical solutions to problems.

**Intuitives** prefer to focus on possibilities and the big picture, easily see patterns, value innovation, and seek creative solutions to problems.

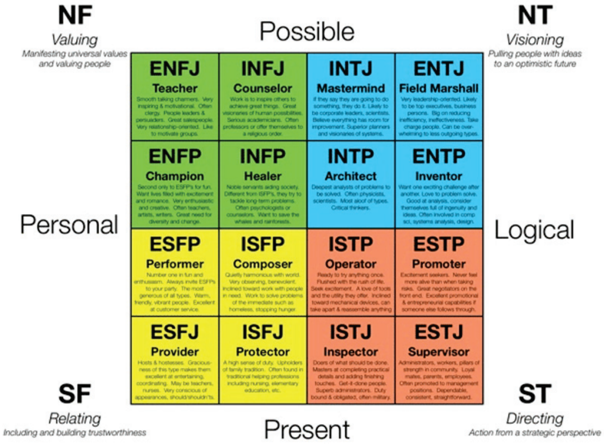
**Judgers** tend to be organized and prepared, like to make and stick to plans, and are comfortable following most rules.

**Perceivers** prefer to keep their options open, like to be able to act spontaneously, and like to be flexible with making plans.

Fig. 1. Personality Types

### B. Personality types in the Myers–Briggs Type Indicator

This project uses the MBTI principle as a guideline to help determine the user's personality based on the following personality dimensions: Introvert (I) and Extrovert (E), Sensation (A) and Intuition (N), Thinking (T) and Feeling (F), Perceiving (P) and Judging (J). The combination of the four types of personality characteristics mentioned above results in sixteen different types of personalities, such as "INFJ" or "ENFP". 4


NF
Valuing
Manifesting universal values and valuing people
Possible
NT
Visioning
Pulling people with ideas to an optimistic future
Personal
Logical
SF
Relating
Including and building trustworthiness
Present
ST
Directing
Action from a strategic perspective

| | | | |
|---|---|---|---|
| **ENFJ**<br>**Teacher**<br>Smooth talking charmers. Very inspiring & motivational. Often clergy. People leaders & persuaders. Great salespeople. Very relationship-oriented. Like to motivate groups. | **INFJ**<br>**Counselor**<br>Work is to inspire others to achieve great things. Great visionaries of human possibilities. Serious academicians. Often professors or offer themselves to a religious order. | **INTJ**<br>**Mastermind**<br>If they say they are going to do something, they do it. Likely to be corporate leaders, scientists. Believe everything has room for improvement. Superior planners and visionaries of systems. | **ENTJ**<br>**Field Marshall**<br>Very leadership-oriented. Likely to be top executives, business persons. Big on reducing inefficiency, ineffectiveness. Take charge people. Can be overwhelming to less outgoing types. |
| **ENFP**<br>**Champion**<br>Second only to ESFP's for fun. Want lives filled with excitement and romance. Very enthusiastic and creative. Often teachers, artists, writers. Great need for diversity and change. | **INFP**<br>**Healer**<br>Noble servants aiding society. Different from ISFP's, they try to tackle long-term problems. Often psychologists or counselors. Want to save the whales and rainforests. | **INTP**<br>**Architect**<br>Deepest analysts of problems to be solved. Often physicists, scientists. Most aloof of types. Critical thinkers. | **ENTP**<br>**Inventor**<br>Want one exciting challenge after another. Love to problem solve. Good at analysis, consider themselves full of ingenuity and ideas. Often involved in comp sci, systems analysis, design. |
| **ESFP**<br>**Performer**<br>Number one in fun and enthusiasm. Always invite ESFPs to your party. The most generous of all types. Warm, friendly, vibrant people. Excellent at customer service. | **ISFP**<br>**Composer**<br>Quietly harmonious with world. Very observing, benevolent. Inclined toward work with people in need. Work to solve problems of the immediate such as homeless, stopping hunger. | **ISTP**<br>**Operator**<br>Ready to try anything once. Flushed with the rush of life. Seek excitement. A love of tools and the utility they offer. Inclined toward mechanical devices, can take apart & reassemble anything | **ESTP**<br>**Promoter**<br>Excitement seekers. Never feel more alive than when taking risks. Great negotiators on the front end. Excellent promotional & entrepreneurial capabilities if someone else follows through. |
| **ESFJ**<br>**Provider**<br>Hosts & hostesses. Graciousness of this type makes them excellent at entertaining, coordinating. May be teachers, nurses. Very conscious of appearances, should/shouldn'ts. | **ISFJ**<br>**Protector**<br>A high sense of duty. Upholders of family tradition. Often found in traditional helping professions including nursing, elementary education, etc. | **ISTJ**<br>**Inspector**<br>Doers of what should be done. Masters at completing practical details and adding finishing touches. Get-it-done people. Superb administrators. Duty bound & obligated, often military. | **ESTJ**<br>**Supervisor**<br>Administrators, workers, pillars of strength in community. Loyal mates, parents, employees. Often promoted to management positions. Dependable, consistent, straightforward. |

Fig. 2. MBTI Personality Types

## IV. Methodology

This section details the chronological steps followed in our analytical study of personality prediction and group formation.

### A. Dataset

Myers Briggs Personality Type Dataset(MBTI) has been collected from Kaggle and it contains approximately 8600 rows of data, in each row, there is a person's top 50 social media posts.

### B. Sentiment Analysis

It is used to examine the emotional tone of the post and for data exploration and analysis using the polarity and subjectivity of posts grouped by the personality types.

*1) Polarity of posts:* The orientation of the stated emotion is determined by the element's sentiment polarity, which determines whether the text communicates the user's positive, negative, or neutral sentiment toward the entity in question.

The average polarity is 0.13 and the overall polarity is quite neutral although ENFJ Fig. 4 depicted the highest polarity.

*2) Subjectivity:* In order to determine a person's feeling, emotions and judgments the average subjectivity of the post is calculated to be 0.54. Thus, the posts have a moderate tone of neither objective nor completely subjective.

### C. Data Cleaning

The dataset acquired contains textual posts by individuals on social media and may contain separators, links and urls and various other symbols and punctuation marks.

The data thus, was cleaned using the following techniques:

- Replacing all ”— — —” (post separators) with ’ ’ in the posts
- Removing all the stop words from the posts
- Applying post stemming
- Conversion of words to lowercase
- Removal of the hyperlinks with ‘URL’
- Removal of digits and punctuation

### *D. Data Preprocessing*

*1) TF-IDF:* In order to transform the textual posts into columns of input data, TF-IDF vectorizer is used. The given Equation: 1 involves the product of two metrics: the number appearances of a word in a document and the word’s inverse document frequency over a collection of documents. This method is used as it is based on the relevance of individual word to a document in a group of documents.

$$W_{i,j} = tf_{i,j} * log(\frac{N}{df_i}) \qquad (1)$$

*2) Truncated SVD:* The dataset contained a large variety of words used in the posts, the feature vector seemed to be quite large, therefore, truncated SVD was used for dimensionality reduction.

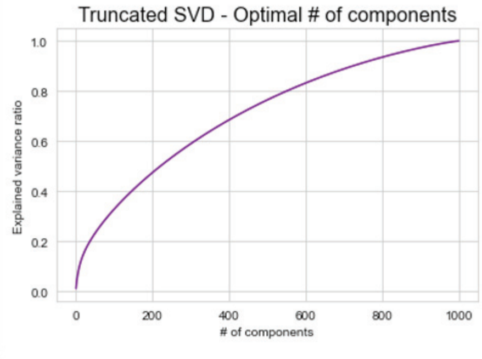


Fig. 3. Truncated SVD for dimensionality reduction

### *E. Training And Testing*

Text Classification is a more essential job in Supervised Machine Learning. Multi class classification and binary class classification models are used for text categorization. The training and validation data sets are distributed in a 7:3 ratio. As a result, 30% of the data is put aside for validation.

*1) Multi Class Classification:* In multi class classification of dataset the algorithms that are taken into consideration are random forest classifier, k - nearest neighbours and one vs rest classifier.

*2) Binary Class Classification:* We have seen in the above classification that the models could not achieve a high enough accuracy and therefore, binary classification on individual axes of I-E (Introversion vs Entroversion), N-S (Intuition vs Sensing), F-T (Feeling vs Thinking) and J-P (Judging vs Perceiving) was done. Further, some performance was improved by handling the imbalance of data using SMOTE technique of oversampling using synthetic data creation. In binary class classification of dataset the algorithms that are taken into consideration are random forest classifier, one vs rest classifier, multi-layered perceptron model and extreme gradient boosting.

*3) Models:*

*a) Random Forest Classifier:* Random Forest Classifiers comprises of a great many number of decision trees, which are trained in such a manner as to avoid the overfitting problem. Also, It uses the bagging technique for constructing the trees.
The following parameters were used for the tuning of this classifier:

- n estimators: 30
- min samples leaf: 50
- oob score: True
- n jobs: -1

*b) K Nearest Neighbours:* The K-NN algorithm is a non parametric classification strategy in statistics that is used for classification and regression. In all cases, the input is the k closest training samples in a data set. It seeks the ”closest” observations in its training data to the observation to predict, then averages or votes on the target values of those training observations to estimate the value for the new data point. Using GridSearchCV, the optimal number of neighbours was discovered to be 15.

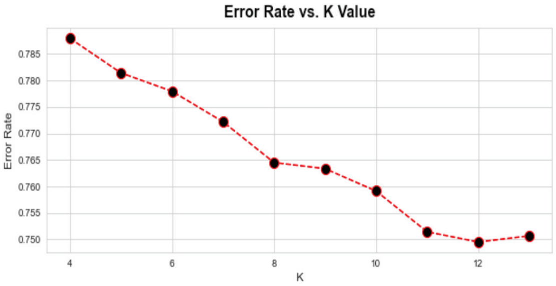


Fig. 4. Error Rate vs K value

*c) One Vs Rest Classifier:* One-vs-rest classification is a multi-class classification method that makes use of binary classification techniques. It is necessary to divide the multi-class dataset into a set of binary classification jobs. A binary classifier is trained for each binary classification task, and predictions are made using the model with the highest confidence. A one versus rest classifier with LinearSVC and random state = 123 is used to classify posts.

*d) Multi-layer Perceptron:* MLP employs a supervised learning approach known as back propagation. The MLP model includes dropout layers to reduce overfitting and randomly switches input units to 0 with a rate frequency at each step during training. A multi-layer perceptron model with two dropout layers and two dense with activation function of relu with a the last dense layer of activation function sigmoid was used for the classification of posts. The loss of sparse categorical crossentropy was used. Activation function of relu is used which is fast than other activation functions and removes the Vanishing Gradient problem, hence providing better prediction accuracy and efficiency

*e) Extreme Gradient Boosting:* XGBoost is a machine learning technique that constructs ensembles using decision trees that are introduced to the ensemble one by one and optimized to correct the prediction mistakes caused by previous models.

$$Gain = LeftLeaf_{similarity} + RightLeaf_{similarity} - Root_{similarity} \tag{2}$$

$$Score = \frac{(\sum_{n}^{i=1} Residual_i)^2}{\sum_{n}^{i=1}[PrevProb_i * (1 - PrevProb_i)] + \lambda} \tag{3}$$

*4) Cluster Formation:* People use social media to express themselves, and their posts may be utilised to help them find the right friend groups and communities. As a consequence, we used the elbow approach to calculate the appropriate number of clusters using the unsupervised K means clustering algorithm, which may then be used to present them with suitable group selections.

## V. Results

The metrics used for comparing the performance of models are accuracy, f1-score and ROC AUC.

$$Accuracy = \frac{TN + TP}{TP + TN + FP + FN} \tag{4}$$

When dealing with uneven data, ROC analysis has no bias in favour of models that perform well on the minority class at the expense of the majority class, which is a very appealing feature. The area under the roc curve can be calculated for all threshold values to offer a single score for a classifier model and this area is termed as ROC AUC (Receiver Operating Characteristic Curve Area Under Curve).

ROC Curve: Plot of False Positive Rate (x) Eq: 6 vs. True Positive Rate (y) Eq: 5.

$$TruePositiveRate = \frac{TP}{TP + FN} \tag{5}$$

$$FalsePositiveRate = \frac{FP}{FP + TN} \tag{6}$$

### A. Multi Class Classification

In order to categorise the data into their 16 potential personality types, the four models of Random forest classifier, K nearest neighbours, One vs Rest classifier, and multi-layer perceptron models were utilised.

For Random Forest classifier all predictions limited to only within four of the sixteen classes: INFJ, INFP, INTJ, and INTP, which are the most abundant classes in the dataset. In K-Nearest Neighbours the model was clearly overfit. For The One vs Rest classifier the predictions were scaled out to all the classes and not just the dominant classes and in Multi Layered Perceptron the model gave the best overall results as compared to other models.

TABLE I
Multi-class Classification Results

| Models | Train Accuracy | Test Accuracy |
|---|---|---|
| Random forest classifier | 41% | 21.1% |
| K Nearest Neighbours | 100%. | 27.19% |
| One vs rest classifier | 44.38% | 36.84% |
| Multi-layer perceptron | 72.46% | 31% |

### B. Binary Classification

The categorization of posts is done on four axes. I vs E, T vs F, N vs S and J vs P.

- The dataset imbalance was handled by using the SMOTE technique of oversampling by artificially constructed data.
- Random forest Classifier, One vs Rest Classifier, Multi layer perceptron model and XGboost were used to predict the four axes individually.
- The binary classification accuracy outperformed the multiclass classification accuracy by a wide margin.
- Although still the predictions tended to favor the more abundant cases.

### C. Cluster Formation

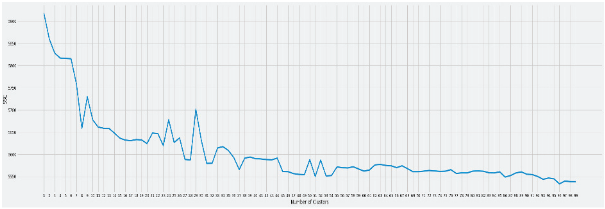


Fig. 5. Sum of Squared Errors(SSE) VS Number of Clusters

The ideal number of clusters for the k means clustering technique is determined using the graph Fig: 5. After evaluating the presented graph, the elbow approach revealed that the ideal number of clusters is 37.

## VI. Conclusion

This study shed light on a variety of machine learning strategies for simplifying the process of finding personality types using the MBTI personality type indicator. Sentiment Analysis is utilized to examine the emotional state of the posts. Natural language processing is employed in the development process, along with additional approaches such as Truncated SVD and SMOTE. Random Forest Classifier, KNN, One vs Rest, MLP and XG-Boost are the models utilised for classification purposes.K-means Clustering is used for finding the best number of groups to have in a network based on their personalities. Python libraries utilised included Pandas, Numpy, Seaborn, Matplotlib, NLTK, Tensorflow, Keras, and Sklearn. The models' accuracy, F1-score, and ROC AUC are evaluated, and

TABLE II
BINARY CLASS CLASSIFICATION RESULTS

| Models | I vs E | N vs S | T vs F | J vs P |
|---|---|---|---|---|
| Extreme Gradient Boost | Train accuracy : 91.6%<br>Test accuracy : 81%<br>f1 score for introversion : 0.46<br>f1 score for extroversion : 0.89<br>ROC accuracy : 93.83% | Train accuracy : 94.64%<br>Test accuracy : 86.0%<br>f1 score for intuition : 0.92<br>f1 score for sensing : 0.27<br>ROC accuracy : 96.76% | Train accuracy : 90.17%<br>Test accuracy : 79.0%<br>f1 score for feeling : 0.81<br>f1 score for thinking : 0.77<br>ROC accuracy : 89.56% | Train accuracy : 87.1%<br>Test accuracy : 73.0%<br>f1 score for judging : 0.62<br>f1 score for perceiving : 0.8<br>ROC accuracy : 86.13% |
| Random forest classifier | Train accuracy : 78.7%<br>Test accuracy : 78.0%<br>f1 score for introversion : 0.0<br>f1 score for extroversion : 0.88<br>ROC accuracy : 86.937% | Train accuracy : 97.9%<br>Test accuracy : 79.1%<br>f1 score for intuition : 0.88<br>f1 score for sensing : 0.27<br>ROC accuracy : 96.089.13% | Train accuracy : 94.0%<br>Test accuracy : 71.6%<br>f1 score for Feeling : 0.74<br>f1 score for Thinking : 0.69<br>ROC accuracy : 74.349% | Train accuracy : 95.8%<br>Test accuracy : 59.0%<br>f1 score for judging : 0.52<br>f1 score for perceiving : 0.64<br>ROC accuracy : 86.13% |
| One vs rest classifier | Train accuracy : 73.3%<br>Test accuracy : 69.6%<br>f1 score for introversion : 0.74<br>f1 score for extroversion : 0.73<br>ROC accuracy : 77.937% | Train accuracy : 75.9%<br>Test accuracy : 71.8%<br>f1 score for intuition : 0.82<br>f1 score for sensing : 0.38<br>ROC accuracy : 96.089% | Train accuracy : 80.0%<br>Test accuracy :77.6%<br>f1 score for Feeling : 0.79<br>f1 score for Thinking : 0.76<br>ROC accuracy : 86.185% | Train accuracy : 69.1%<br>Test accuracy : 64.5%<br>f1 score for judging : 0.58<br>f1 score for perceiving : 0.69<br>ROC accuracy : 72.042% |
| Multi-layer perceptron model | Train accuracy : 98%<br>Test accuracy : 76.3%<br>f1 score for introversion : 0.86<br>f1 score for extroversion : 0.21 | Train accuracy : 98.8%<br>Test accuracy : 74.9%<br>f1 score for intuition : 0.92<br>f1 score for sensing : 0.07 | Train accuracy : 91.11%<br>Test accuracy : 68.92%<br>f1 score for Feeling : 0.75<br>f1 score for Thinking : 0.59 | Train accuracy : 85.9%<br>Test accuracy : 53.28%<br>f1 score for Judging : 0.51<br>f1 score for Perceiving : 0.56 |

the outcomes are compared. This work can automate the determination of personality type and related thought processes.

## VII. Future Work

Further work can be done in future to improve the efficiency and try out other techniques to improve the study on the dataset. Some of the potential development are as follows:

- Improving the imbalances in the data and incorporating data from other personality types with more of sensing quality
- Use of penalized models to handle the imbalance. If the model makes a classification error on the minority class during training, it will suffer additional costs. The model is able to favour the minority group because of these penalties.
- Using word2vec and RNN models for classification.
- Exploration of algorithms for Community detection for the dataset to detect and extract the groups with similar properties.

## References

[1] Mohammad Hossein Amirhosseini and Hassan Kazemian. Machine learning approach to personality type prediction based on the myers–briggs type indicator®. *Multimodal Technologies and Interaction*, 4(1):9, 2020.

[2] Shristi Chaudhary, Ritu Singh, Syed Tausif Hasan, and Ms Inderpreet Kaur. A comparative study of different classifiers for myers-brigg personality prediction model. *Linguistic analysis*, page 21, 2013.

[3] Paolo Fornacciari, Monica Mordonini, Agostino Poggi, Laura Sani, and Michele Tomaiuolo. A holistic system for troll detection on twitter. *Computers in Human Behavior*, 89:258–268, 2018.

[4] Prantik Howlader, Kuntal Kumar Pal, Alfredo Cuzzocrea, and SD Madhu Kumar. Predicting facebook-users' personality based on status and linguistic features via flexible regression analysis techniques. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*, pages 339–345, 2018.

[5] Alam Sher Khan, Hussain Ahmad, Muhammad Zubair Asghar, Furqan Khan Saddozai, Areeba Arif, and Hassan Ali Khalid. Personality classification from online text using machine learning approach. *International Journal of Advanced Computer Science and Applications*, 11(3), 2020.

[6] Aditi V Kunte and Suja Panicker. Using textual data for personality prediction: a machine learning approach. In *2019 4th international conference on information systems and computer networks (ISCON)*, pages 529–533. IEEE, 2019.

[7] Tuan Tran, Dong Nguyeny, Anh Nguyeny, and Erik Golenz. Sentiment analysis of emoji-based reactions on marijuana-related topical posts on facebook. In *IEEE International Conference on Communications (ICC)*, 2018.