Predicting Personality Types using Machine Learning and the Myers-Briggs Inventory

Indrajeet Kumar,
Asst. Professor, Department of Comp. Sc. & Info. Tech., Graphic Era Hill University, Dehradun, Uttarakhand India 248002

Abstract
It is possible to study the reviews-based dataset. The validity of this paradigm was demonstrated by our data analysis. The data taken from the Twitter reviews (Myer-Briggs) dataset is subjected to behaviour analysis. To determine the user's remark, the data is evaluated. We want to use data-driven marketing technologies including supervised machine learning models, natural language processing, along with information visualisation. Logistic Regression is one of the classification techniques on which the system was constructed. The current issues with each topic are examined before the most recent fixes are provided and debated. The findings from the experiment demonstrate the accuracy, precision, recall, and F1 score. Following that, we can use the Twitter API to forecast the personality. It illustrates how openness is compared.

Keywords: Machine Learning, Natural Language Processing, Personality Prediction, Myer-Briggs, Behaviour Analysis.

DOI Number: 10.48047/nq.2021.19.6.NQ21093 NeuroQuantology2021;19(6): 236-241

1. INTRODUCTION

Personality traits refer to a person's cognitive patterns, mannerisms, emotional responses, and other attributes. The input data is in.csv file format. The Myer-Briggs dataset is used to gather the inputs. The pre-processing procedure must then be carried out. We must employ the Natural Language Processing method in that phase. The second step is to employ a classification technique, such logistic regression. The findings then demonstrate performance analysis using metrics including accuracy, precision, recall, and F1-score as well as data visualisation using bar and pie graphs. We provide a supervised learning method to calculate five personality characteristics in this work. According to the Big Five concept, there are five fundamental personality qualities.

According to the Big Five concept, there are five fundamental personality qualities. The quality of openness has a distribution that is generally normal, with the exception of a tiny fraction of persons who score extremely high or low on it, with the majority of people scoring somewhere in the middle. People with low scores are commonly thought to be resistant to new experiences. They frequently display attitudes and practises that are considered to be conventional and traditional. They are more likely to have a narrower range of interests and to choose well-established habits above trying out new things. Openness and creativity, intellect, and knowledge have mildly beneficial correlations. Similar to how absorption is a psychological attribute, openness also has a moderate association to individual variations in hypnotic susceptibility.

Being conscientious is taking one's societal responsibilities seriously and having the desire to complete one's work in an accurate manner. People who pay attention to detail tend to be efficient and well-organized, as opposed to being sloppy and disorganised. They are typically reliable, conduct responsibly, fulfil their responsibilities, and work hard to achieve their goals. In addition to this, they act in a planned fashion as opposed to acting impulsively. Being agreeable is a superordinate feature, which implies that it significantly clusters with other personality subtraits. This means that being agreeable is a desirable trait to possess.Lower-level characteristics or components that go into making up agreeableness include trust, candour, altruism, obedience, modesty, and open-mindedness.

Because extraversion and introversion are often considered to coexist on the same continuum, it follows that having a high level of one implies having a low level of the other. Jung proposes an alternative viewpoint in which he asserts that

everyone possesses both an extraverted and an introverted side, with one of these sides being more predominant than the other in any given individual. There are a lot of different personality theory models that contain neuroticism as a trait, but there is a lot of disagreement over how to define it. Others characterise it as versus emotional stability and positivity, or a good adjustment, how emotionally unstable and negative you are. Some define it as a predisposition for fast arousal when prompted and sluggish relaxation following arousal, especially in relation to negative emotional arousal. Social networking sites are also utilised as venues for people to share their ideas, feelings, and essays.

They are frequently presented in front of a big audience in a public setting. For instance, one of the most widely used platforms is Twitter, where tweets are eventually used to characterise the people who write them or share the tweets of others. To guide our research, we developed the following question: Can important features of an individual be gleaned from tweets to what extent? Language psychology has actively researched the topic of word use profiling. In reality, language psychology's fundamental premise is that each person's choice of words reflects their personality. For instance, researchers found that people's speech patterns correlate with both physical as well as mental health issues.

Computer technology has made it possible to look for psychological patterns or people's underlying emotions that could be reflected in the language they used. (The hypothesis that people's personalities are implicitly stored in the words they use to make sentences has been supported by several psychological and linguistic investigations.) Using solely what a person tweets about their ideas, we offer a supervised learning strategy to compute five personality characteristics in this study. The method divides tweets into tokens, learns word vector representations as embeddings, and then feeds supervised learner classifier with these representations. By calculating the mean squared error of the learnt paradigm employingainternational standard of Facebook status updates, we show the efficacy of the method.

We also used a benchmark created by 24 panellists who participated in a cutting-edge psychological survey to test the transfer learning potential and viability of this approach, and we saw good alignment between the model's predictions of Twitter messages and the personality characteristics discovered by the study. The remainder of this essay is organised as follows: we provide the 'Five Factor Model' and compare our strategy to cutting-edge learning algorithms. We outline the methodology, and in Section 6 we give the experimental findings during the process of analysing personality features in tweets, we made. We provide the norm from Twitter as well as the gold standard that was utilised to train the learning model in this study.

2. LITERATURE SURVEY

Author demonstrate how readily available digital behaviour records, such as Facebook Likes may be employed to forecast a number of very sensitive personal traits automatically and correctly, such as Age, gender, parental separation, usage of addictive substances, IQ, happiness, sexual orientation, ethnicity, as well as political and religious beliefs. The study's findings are based on a dataset of more than 58,000 participants who gave their Facebook Likes, in-depth demographic profiles, and the outcomes from multiple psychometric tests. To prepare the Likes data for logistic/linear regression to predict individual psych demographic characteristics from Likes, the suggested methodology employs dimensionality reduction [1].

Thirty percent of persons with bipolar disorder will commit suicide, and seven out of ten people with the illness receive incorrect diagnoses at first. One of the essential elements for preventing the complete development of the condition is recognising the early stages of the disorder. The author's goal in this work is to use social media data to create prediction models that make use of psychological and phonological characteristics to ascertain the time of bipolar disorder's beginning and offer details on its prodrome. By utilising a cutting-edge data collecting technique called Time-specific Subconscious Crowdsourcing, this study is able to achieve these discoveries. This method assists in gathering a trustworthy dataset that supports diagnostic data from bipolar illness sufferers. The experimental findings show that the suggested models might make a significant contribution to routine bipolar disorder evaluations, which are crucial in the primary care context.The main drawback of the system is its inability to predict personality traits properly [2].

This paper examines how the way people use language can be used to differentiate individuals. The authors argue that language is an important indicator of psychological and social characteristics, and that studying linguistic patterns can provide insight into a person's emotional state, personality, and social behavior. Pennebaker and King's research suggests that there are significant differences in linguistic style between individuals, and that these differences are related to various psychological and social factors. Overall, Pennebaker and King's article makes a significant contribution to the field of psychology by demonstrating the importance of linguistic style as an individual difference variable. It highlights the

value of language as a tool for understanding human behavioralso offers insights into the complex interplay between language and psychology [3].

The study by Argamon et al. (2005) aimed to investigate the relationship between word usage and personality types. The authors hypothesized that certain words may be more commonly used by individuals with certain personality traits. To test their hypothesis, the researchers analyzed a collection of writings produced by university students along with used linguistic analysis software to identify the frequency of certain words. The authors found that individuals who scored high on the personality trait of extraversion were more likely to use words related to social activities and emotions, while those with higher neuroticism scores were more prone to employ terms that expressed negative emotions. Additionally, individuals who scored high on agreeableness used more words related to social relationships, while those who scored high on openness to experience used more words related to intellectual pursuits.Considering these drawbacks, the study offers useful understandings of the connection between language usage and personality traits. Future research could build on this study by using larger and more diverse samples and incorporating more objective measures of personality traits. Overall, the study contributes to our understanding of the links between language and personality and highlights the potential of natural language processing techniques in psychological research [4].

This paper presents a seminal work pertaining to industrial-organizational psychology. The goal of the study was to determine how the Big Five personality traits as well as job performance, using a meta-analytic approach. The results revealed a clear and consistent relationship between the Big Five personality dimensions and job performance. Conscientiousness discovered to be the most potent predictor of job performance, followed by emotional stability (low neuroticism) and agreeableness. Extraversion and openness to experience had moderate relationships with job performance. Barrick and Mount's study is significant for several reasons, such as providing evidence of the validity of the Big Five personality paradigm as a predictor of job performance, having practical implications for organizations seeking to improve their selection and hiring practices, and stimulating further research on how personality and job success are related. Overall, the study remains a crucial contribution to the field of industrial-organizational psychology [5].

This article provides a comprehensive review of empirical literature on the Five-Factor Model of personality and personality disorders. The findings of the research demonstrated that the FFM has a significant relationship with personality disorders, with the highest correlations found between Neuroticism and Borderline, Dependent, and Avoidant personality disorders, Extraversion and Histrionic and Narcissistic personality disorders, and Openness to Experience and Schizotypal personality disorder. The authors concluded that the FFM could be used as a reliable diagnostic tool for personality disorders, but cautioned that caution should be exercised in its use. Further research is needed to examine the FFM's utility in diagnosing specific personality disorders and its applicability across diverse populations [6].

3. PROPOSED SYSYTEM

Social media platforms are currently the biggest repositories of personal data as they continually record users' behaviours, interactions, and interests in music, movies, as well asshopping.We now put forth a supervised learning strategy to figure out five personality characteristics only based on the content of a person's tweets about their views. The method divides tweets into tokens, learns word vector representations as embeddings, and then feeds the learned word vector representations to a supervised learner classifier. We employed logistic regression and support vector machines in this study. Low Accuracy in addition to its Major flaws in the current method include the comparison graph not being displayed.

The data is processed using the Myer-Briggs dataset of twitter reviews. Logistic Regression is one of the classification techniques on which the system was constructed. OCEAN from the Twitter API allows us to anticipate a person's personality. The findings from the experiment demonstrate the accuracy, precision, recall, and F1 score. The experiment enhanced the precision. And also displays the graphs for visualisation. We can now forecast the personality features. And show how Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism compare to one another.

The following benefits of the suggested strategy are listed:

- Increasing precision.
- Using api, predict a person's personality via twitter.
- Show the graphs used for visualisation.

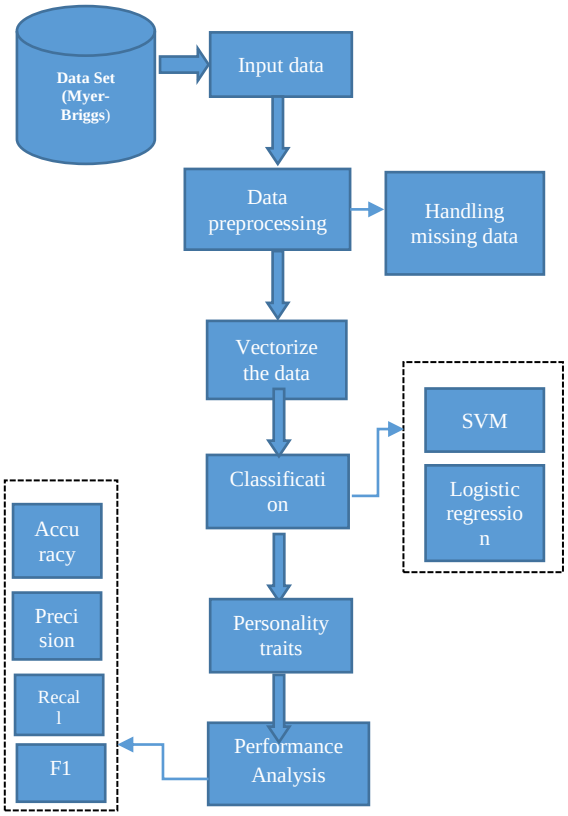

Fig 1: Flow Diagram

The next section provides an explanation of the many phases that are involved in putting the suggested technique into practise:

1. Data Selection and Loading

The process of choosing the data for analytical detection is known as data selection. The Myer-Briggs dataset is utilised in this research. The data collection containing the comments' information.

2. Data Preprocessing

Eliminating superfluous information from a dataset is an important step in the pre-processing of data. Throughout the entirety of this process, missing values and Nan values will be replaced by the value 0 in order to eliminate missing data. The data have been cleaned of any errors, missing values, and duplicates that may have been present. Encoding Information that can be organised into categories: Variables that have a predetermined set of label values are referred to as categorical data. that virtually all algorithms used in machine learning require both numerical input and output variables in order to function properly.

3. Splitting Dataset Into Train And Test Data

The method of dividing data that is accessible into two halves, known as "data splitting," is typically done to satisfy the needs of a cross-validator.

A portion of the data is utilised in the construction of a prediction model, whereas a distinct quantity is utilised in the assessment of the performance of the model. The separation of the data into training and testing sets is an essential action in the process of analysing data mining algorithms. When you divide a data set into a training set and a testing set, the majority of the data is often utilised for training, while just a smaller portion of the data is used for testing.

4. Classification

A non-linear transform is applied to the output of the linear technique known as logistic regression. It does assume that there is a linear connection between the variables that are being input and those that are being output. The application of data modifications to your input variables that more clearly expose this linear link may result in the production of a model that is more accurate.

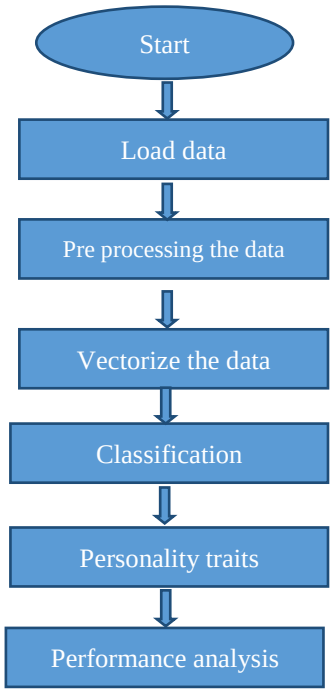

Fig 2: System Architecture

5. Prediction

It involves extrapolating behavioural analysis from a dataset. By optimising the total prediction outcomes, this project will successfully forecast the data from the dataset.

6. Result Generation

On the basis of the overall categorization and forecast, the Final Result will be created. The effectiveness of this suggested strategy is assessed using some metrics, such as,

Accuracy is another name for the classifier's aptitude to divide things up. It makes an accurate

prediction of the class label, and predictor accuracy is defined as the degree to which a particular predictor can reliably forecast the value of an attribute based on newly collected data.

$$AC= (TP+TN)/ (TP+TN+FP+FN)$$

The term "precision" refers to the ratio of the number of genuine positives to the total number of positive results, including both true and false positives.

$$Precision=TP/ (TP+FP)$$

The percentage of information that can be recalled is determined by taking the total number of correct outcomes and dividing that by the total number of expected results. In binary classification, what is known as recall is actually referred to as sensitivity. It is possible to think of it as the possibility that the query will obtain a document that is relevant to the inquiry.

$$Recall=TP/ (TP+FN)$$

The F measure, commonly referred to as the F1 score or simply the F score, is a metric that determines how accurate a test is by calculating the weighted harmonic mean of the precision and recall scores associated with the test.

$$F\text{-}measure=2TP/ (2TP+FP+FN)$$

4. RESULTS

Our data study showed the validity of a dataset based on reviews. Using supervised machine learning models, natural language processing, and information visualisation, the data from the Twitter reviews (Myer-Briggs) dataset is exposed to behaviour analysis and appraised. One of the categorization methods used to build the system is logistic regression. The experiment's results show the F1 score, recall, accuracy, and precision. The Myer-Briggs dataset of Twitter reviews is used to process the data.

One of the categorization methods used to build the system is logistic regression. The results of the experiments demonstrate the greatest levels of accuracy, precision, and F1-score were reached by the suggested method, outperforming machine learning approaches. The outcomes will compare various personality traits.

```
(6072,) (6072,) (2603,) (2603,)
.....LOGISTIC REGRESSION....

ACCURACY: 60.54552439492893

              precision    recall  f1-score   support

           0       0.61      0.66      0.63       453
           1       0.55      0.56      0.56       194
           2       0.60      0.70      0.65       388
           3       0.61      0.61      0.61       349
           4       0.60      0.39      0.47        71
           5       0.65      0.27      0.38        64
           6       0.63      0.76      0.69       532
           7       0.58      0.54      0.56       189
           8       0.51      0.37      0.43        78
```

IPython console | History log

Fig 3: Performance Measure

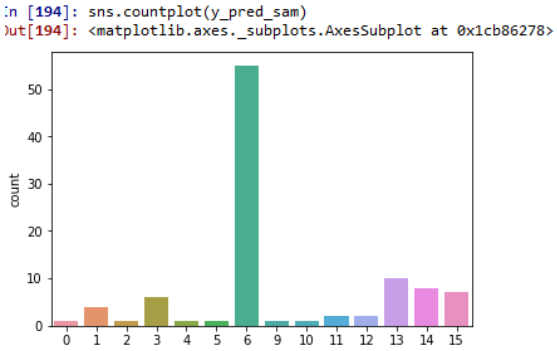

Fig 4: Performance Analysis

5. CONCLUSION

In this study, we developed a method for computing personality characteristics using supervised learning. We want to use data-driven marketing technologies including supervised machine learning models, natural language processing, and data visualisation. Support vector machine and logistic regression are two classification techniques on which the system was constructed. The experimental findings show that the suggested technique performed better than machine learning algorithms and attained the maximum accuracy, precision, and F1-score.The outcome will show how personality characteristics compare.

REFERENCE

[1] Kosinski, M.; Stillwell, D.; Graepel, T. Private traits and attributes are predictable from digital records of human behavior. Proc. Natl. Acad. Sci. USA 2013, 110, 5802–5805.

[2] Gottschalk, L.A.; Gleser, G.C. The Measurement of Psychological States through the Content Analysis of Verbal Behavior; University of California Press: Oxford, UK, 1969.

[3] Pennebaker, J.; King, L.A. Linguistic Styles: Language Use as an Individual Difference. Personal. Soc. Psychol. 1999, 77, 1296–1312.

[4] Argamon, S.; Dhawle, S.; Koppel, M.; Pennebaker, J. Lexical predictors of personality type. In Proceedings of the 2005 Joint Annual Meeting of the Interface and the Classification Society of North America, Cincinnati, OH, USA, 24–28 July 2005.

[5] Barrick, M.; Mount, M. The Big Five personality dimensions and job performance: A meta-analysis. Pers. Psychol. 1991, 44, 1–26. [CrossRef]

[6] Saulsman, L.; Page, A. The five-factor model and personality disorder empirical literature: A meta-analytic review. Clin. Psychol. Rev. 2004, 23, 1055–1085

[7]Gosling, S.D.; Augustine, A.A.; Vazire, S.; Holtzman, N.; Gaddis, S. Manifestations of Personality in Online Social Networks: Self-

Reported Facebook-Related Behaviors and Observable Profile Information. Cyberpsychol. Behav. Soc. Netw. 2011, 14, 483–488.