JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

ANALYSIS OF HUMAN TRAITS USING MACHINE LEARNING

Mr. Padmalaya Archith. Anurag Group of Institutions (IT), Hyderabad, India
Mr. Sirigiri Shiva Sai. Anurag Group of Institutions (IT), Hyderabad, India
Ms. Snigdha Neela. Anurag Group of Institutions (IT), Hyderabad, India
Mrs. N. Naga Lakshmi. Anurag Group of Institutions (IT), Assistant Professor, Hyderabad, India

Abstract - Personality is a crucial factor since it distinguishes one person from another. Personality prediction has numerous uses in the real world. The main goal of this project is to collect the user's textual data and apply a trained machine learning model to predict his four personality traits: Introversion vs Extroversion, Sensing vs Intuition, thinking vs Feeling, and Judging vs Perceiving. The main goal is to create an application that allows users to answer questions that are then processed and evaluated to determine his personality features. The result is a four-character string in which each character represents a different personality feature; a total of sixteen personality types are possible. The text is classified using the machine learning algorithm Random Forest Classifier, which produces four personality traits. Natural Language Processing (NLP) approaches with the help of NLTK libraries will be used to process and categorize enormous amounts of textual data. Hyper parameter tuning and cross fold validation are used to improve the performance of the model.

1. INTRODUCTION

1.1 Introduction

Personality is what sets people apart from one another, thus it's a vital factor to consider. Personality is an important part of human existence. The study of personality falls under the umbrella of psychological research. Personality is made up of aspects that change with time, such as a person's thoughts, feelings, and conduct. People are divided into different groups of personality types, hence personality prediction is considered as a classification problem in computer science. Different sorts of personality classifications can be determined using a variety of psychological tests. MBTI, Big Five, and DISC are all popular personality assessments. One of the most well-known and extensively used personality tests or descriptions is the Myers-Briggs Type Indicator (MBTI). It defines how individuals interact and behave. With four binary categories and 16 total types, they interact with the world around them. Introversion vs. Extroversion, Sensing vs. Intuition, thinking vs. Feeling, judging vs. Perceiving are the differences. Understanding personality features may be extremely beneficial since it allows users to learn why people behave in various ways, identify areas where they can develop, and locate others who share similar personality qualities. The project's main goal is to create an application in which users answer a few questions, which are then evaluated and personality traits are generated. The result is a string of four characters, each of which represents a personality trait, resulting in a total of 16 personality types. The MBTI personality type of each individual is unique. The sum of their four types for each of the four categories, using the bolded distinguishing letter for each. For example, someone who gets most of their energy from being around other people (E) trusts their instincts and uses intuition to make decisions. N interprets global information, T thinks rationally about their judgments, and L lives life in a The personality type ENTJ has a meticulously planned style (J) rather than a spontaneous one.

1.2 Objective

The Random Forest Classifier model will be used to predict personality traits using data from a questionnaire in which the user is asked to answer a series of questions. In the literature, there were two approaches to the problem. The first method was to employ a 16-class metaclassifier, while the second method was to utilize four binary classifiers. The accuracy of the latter strategy was determined to be better since it appeared to be more efficient. AdaBoost, Multinomial Nave Bayes, and LDA were the algorithms employed in the study. The maximum accuracy achieved was 73 percent, according to the results. The literature review concludes that the ensemble is the best option. Because algorithms are more efficient and tuned, they deliver superior performance. As a result, the project's core model is the Random Forest Classifier, which is an ensemble model based on trees.

Problem Formulation:

The development approach will employ the Natural Language Processing Toolkit (NLTK) and the Random Forest Classifier, a supervised machine learning technique. NLTK is a robust natural language processing toolbox for Python developers working with human language data. To analyse the distribution of type indicators in the dataset, four separate categories for type indicators can be developed. Introversion (I)/Extroversion (E) is the first category, while Intuition (N)/Sensing (S) is the second, Thinking (T)/Feeling (F) is the third, and Judging (J)/Perceiving (P) is the fourth (P). As a result, one letter will be returned for each category, resulting in four letters representing one of the 16 MBTI personality types. For example, if the first category returns I, the second category returns N, the third category returns T, and the fourth category returns J, the personality type in question is INTJ. Because the dataset contains raw data, it cannot be utilised to train the machine learning model without first being pre-processed. There are two different forms of classification in machine learning. The goal of the first type, which is based on a set of observations, is to determine whether the data contains classes or clusters. A certain number of classes may exist in the second type, and the goal is to construct a rule or collection of rules to classify a new observation into one of the existing classes.

2. Literature Survey

A Machine Learning Approach to Predicting Personality Using Textual Data

Personality is a crucial factor since it distinguishes people from one another. Personality prediction is a growing field of study. Predicting personality using data from social media is a promising strategy because it eliminates the need for users to fill out surveys, saving time and enhancing credibility. Personality knowledge is thus an intriguing subject for scholars to explore. Personality prediction has numerous uses in the real world. The use of social media is growing every day. Every day, massive amounts of written and visual material are added to the internet. Over Twitter standard, current work focuses on Linear Discriminate Analysis, Multinomial Naive Bayes, and AdaBoost.

MBTI Personality Type Prediction Survey Analysis of Machine Learning Methods for Natural Language Processing

We compared the results of various natural language processing approaches and machine learning techniques in classifying someone's Myers-Briggs personality type based on one of their social media postings. We set out to determine a person's MBTI personality type based on a social media post. Our technology takes a text excerpt as input and generates a predicted MBTI personality label (e.g. ENTJ). We'll look at a range of strategies for this problem, including traditional Supervised Learning and the usefulness of deep learning with actively learned word embeddings. After that, we compare and analyze their error and accuracy results.

Using Traditional and Deep Learning, determine your personality type using the Myers-Briggs Type Indicator and text posting style.

Individual distinctions in thinking, feeling, and behaviour patterns can be reflected as personality. This paper uses a combination of machine learning techniques, such as Naive Bayes, Support Vector Machines, and Recurrent Neural Networks, to predict people's personalities from text using the Myers-Briggs Type Indicator (MBTI). Furthermore, the learning process is guided by CRISP-DM, which stands for Cross-Industry Standard Process for Data Mining. Because CRISP-DM is an iterative development method, we combined it with agile methodology, which is a rapid iterative software development method, to shorten the development cycle..

3. OVERVIEW OF THE SYSTEM

3.1 Existing System

To predict personality types, existing solutions used Naive Bayes, SVM, and LDA as classifiers, and some of them used a multi-class classification technique. But their accuracy, performance and speed are quite low.

3.2 Proposed System

Module for the front-end HTML, CSS, and Bootstrap4 are used to create the Front-End Module. Users can access the software through the front-end module. The project is described on a page on the front end. A form with questions follows, followed by two options. The users' answers for all of the questions is combined to create a single input text on which the model is run to determine personality traits. The findings are displayed on the results page, which includes the user's personality as well as a few characteristics of that personality type.

Module of Integration The input from the front-end module is received and sent to the trained model in this module. The front-end module displays the output generated by the trained model. Flasks are used to construct it. The models are saved as pickle objects after training and are loaded using the flask framework and used to predict the model. Flask Jinja templates are used to display the generated results.

Advantages of Proposed System

- ✓ Automates process of personality prediction detection
- ✓ Previous datasets are used to training and testing.
- ✓ Accuracy of model is improved compared to existing methods.

3.3 Proposed System Design

In this project work, we used five modules and each module has own functions, such as:

1. Data collection
2. Data preprocessing
3. Testing training
4. Initializing Multiple Algorithms
5. Predict data

3.3.1 *Data Collection*

The "MBTI" dataset, which contains 8675 rows of post data and is publicly available on "Kaggle". Personality type based on "MBTI" and personal social networking postings from "personalitycafe.com" are the two columns in each row. Users must first fill out a questionnaire to identify their "MBTI" personality type. Then they can talk or forum with another person in public. because each user has fifty posts at their disposal In total, there are 430,000 data points.

3.3.2 Data Preprocessing

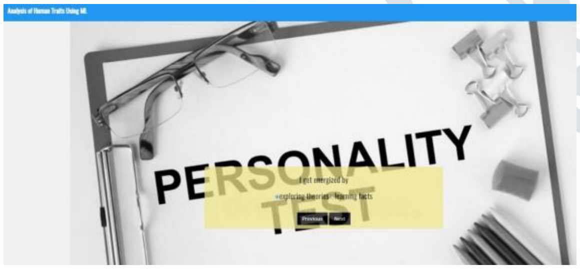
We added four more columns to the datasets that were categorized based on respondents' replies to the four dimensions of "MBTI": "Intuition (N) - Sensation (S)", "Feeling (F) - Thinking (T)", and "Perception (P) - Judgment (J)" are all examples of Introversion. The procedure's goal is to make "Naive Bayes," "Support Vector Machines," and "Recurrent Neutral Networks" more accurate. With "pd.get dummies ()," we created four columns on the four dimensions of "Recurrent Neutral Networks," and used those variables as one-hot encodings. values. Removal of specific words and characters Since the post data originates from "personalitycafe.com," a chat/forum where people only converse with written language. Because we intended our model to be universal, we eliminated several data points that had links to websites. in the language of English In addition, we used "Python's NLTK (Natural Language Toolkit)" to eliminate "stop words" from the text. The Python package "NLTK (Natural Language Toolkit)" is used to process natural language. Lemmatization We'll look at how the root word's inflected forms are turned into dictionary forms. "nltk. stem.wordNetLemmatizer" was used to lemmatize the text. This allows us to capitalize on the fact that inflections still have a common meaning. Tokenization for "Naive Bayes" and SVM Classifiers Using a "NLTK word tokenizer," we tokenized the words that had been altered during the lemmatization process. That is, the common term is broken down into small text fragments. Then we use a "Bag of the Month" to modify the wording to frequency. To investigate the relevance of key-words to documents in the corpus, researchers used "Term Frequency-Inverse Document Frequency (TF-IDF)" and "Term Frequency-Inverse Document Frequency (TF-IDF)" [28,37]. "Recurrent Neutral Networks" Tokenization Using a "Keras word tokenizer," we tokenized the words that had been altered during the lemmatization process. That is, the popular word will shift to position 1, position 2, and so on until there are 74,870 words in total. Any remaining words will now be represented as lists of numbers with a vocabulary ranging from 1 to 74,870. The texts are then converted into sequences with 72,000 number words and a maximum length of 200.

3.3.3 *Testing Training*

The dataset was divided into two parts: training and testing, to determine the accuracy of the "MBTI" personality model. We separated 75 percent of the data for training and 25 percent for testing using the "train-test split" method in the "Scikit-learn" package. The testing dataset is a collection of previously unknown data used simply to assess the effectiveness of a certain classifier.

3.3.4 Initializing Multiple Algorithms

In this stage machine learning algorithms are initialized and train values are given to algorithm by this information algorithm will know what are features and what are labels. Then data is modeled and stored as pickle file in the system which can be used for prediction.


Analysis of Human Traits Using ML

I get energized by

exploring theories · learning facts

Data set is trained with multiple algorithms and accuracy of each model is calculated and best model is used for prediction.

3.3.5 Predict data

In this stage new data is taken as input and trained models are loaded using pickle and then values are preprocessed and passed to predict function to find out result which is showed on web application

4 Architecture

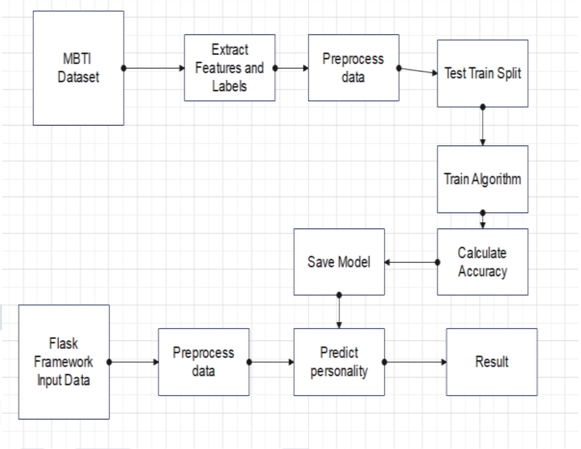
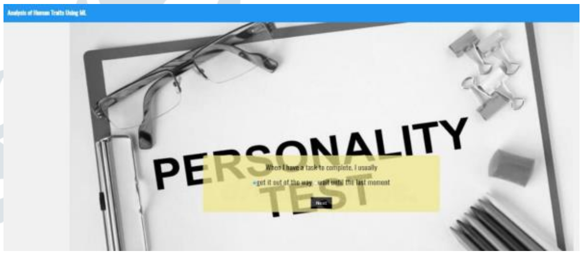


Fig 1: Frame work of Personality test

Above architecture diagram shows three stages of data flow form one module to another module. Back-end storage and machine learning model to train vitamin dataset

5 RESULTS SCREEN SHOTS

Home Page:

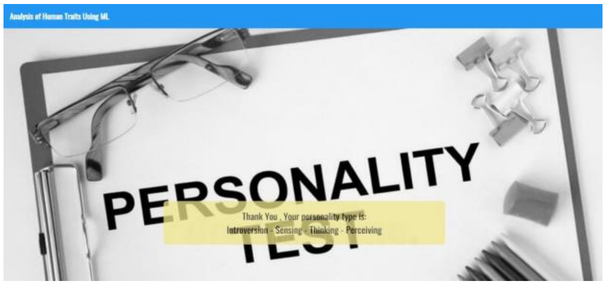

Analysis of Human Traits Using ML

When I have a task to complete, I usually

get it out of the way · wait until the last moment

Answer questions Page:

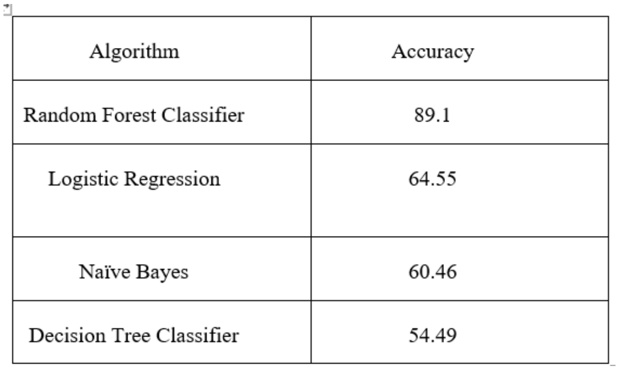
Personality prediction page:


Analysis of Human Traits Using ML

Thank You . Your personality type is:

Introversion - Sensing - Thinking - Perceiving

Accuracy:



Algorithm	Accuracy
Random Forest Classifier	89.1
Logistic Regression	64.55
Naïve Bayes	60.46
Decision Tree Classifier	54.49

7. CONCLUSION

✓ The data is processed using the NLTK package, which includes an integrated list of stop words in the English language. The http links, symbols, and numbers are removed using regular expressions. The best result was achieved by the suggested system for personality prediction utilizing the Random Forest Classifier. The logistic regression model, the Nave Bayes Classifier, and the Decision Tree Classifier all fared worse than the Random Forest Classifier. After hyper parameter adjustment, the Random Forest Classifier's performance improved, resulting in a greater accuracy difference between the Random Forest Classifier and the other models. The dataset is very unbalanced, limiting the system's performance. The system foresees Because the findings are solely based on the responses submitted by the user at the moment, the user's mental state when answering the questions has a stronger influence on the outcome.

Future Enhancement

✓ In the future, the proposed system could be used in real-time applications such as career advice, where one can get advice on what are the best career choices for a person with a particular personality trait, such as an extrovert might be good at jobs requiring a lot of communication, whereas an introvert might be comfortable with jobs requiring minimal communication, and movie/music recommendations,

8. References

- [1] A. V. Kunte and S. Panicker, "Using textual data for Personality Prediction: A Machine Learning Approach," 2019 4th International Conference on Information Systems and Computer Networks (ISCON), 2019, pp. 529-533, doi: 10.1109/ISCON47742.2019.9036220.
- [2] Brandon Cui, Calvin Qi, "Survey Analysis of Machine Learning Methods for Natural Language Processing for MBTI Personality Type Prediction", Stanford,2018
- [3] Hernandez, Rayne, Knight, Ian Scott, "Predicting Myers-Briggs Type Indicator with text classification",31st Conference on Neural Information Processing System, USA,2017
- [4] B. Y. Pratama and R. Sarno, "Personality classification based on Twitter text using Naive Bayes, KNN and SVM," 2015 International Conference on Data and Software Engineering (ICoDSE), 2015, pp. 170-174, doi: 10.1109/ICODSE.2015.7436992
- [5] Shristhi Chaudary, Ritu Singh, Syed Tausif Hasan and Ms. Inderdeep Kaur, A Comparative Study of Different Classifiers for Myers-Brigg Personality Prediction Model, International Research Journal of Engineering and Technology (IRJET),2018
- [6] J. Golbeck, C. Robles, M. Edmondson and K. Turner, "Predicting Personality from Twitter," 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing, 2011, pp. 149-156, doi: 10.1109/PASSAT/SocialCom.2011.33.
- [7] Tommy Tandera, Hendro Derwin Suhartono, Rini Wongso and Yen Lina Prasetio, "Personality Prediction System from Facebook Users", 2nd International Conference on Computer Science and Computational Intelligence, 13–14 October 2017