PERSONALITY PREDICTION PROJECT

PROJECT SYNOPSIS

OF MAJOR PROJECT

BACHELOR OF TECHNOLOGY

COMPUTER SCIENCE ENGINEERING

SUBMITTED BY

ROSHIN JOHN (1800911)

SANDEEP KUMAR (1800913)

TARANDEEP SINGH (1800926)

TARANPREET KAUR (1800927)

GUIDED BY

PROF. KHAYATI MARWAH

BABA BANDA SINGH BAHADUR ENGINEERING COLLEGE

FATEHGARH SAHIB

February (2021)

Table of Contents

INTRODUCTION-

Personality is defined as the set of different Characteristics such as behaviour or emotions as a result of environmental or biological factors. It reflects the persons differences in persons thinking, behaviour and feelings. Personality traits are continuous in nature as they reflect high and low of specific traits in a person on continuous trait rather than showcasing distinct personality. The term personality originally came from the Latin word persona that means mask. There are three criteria that are used to characterize personality traits: consistency along the situations, stability on basis of time and Individual differences that means Different Individuals have different behaviours. The field of study that studies the human personality and its variations among individual and group of people is called Personality Psychology.

With advent of technology the use of Social Networking sites has increased. People use as a platform to express and share their feelings, expectations, experiences etc. Along with this user often share their personal information such as profession, likes and dislikes etc. This extracted data can give businesses opportunity to connect with their customers, understand their needs and thus improve the quality of service or product accordingly. It is used to find the patterns in connectivity, how they are connected and similarity among them. Sentiment analysis is also done on social media data to understand the emotions positive and negative both on some topic by users.

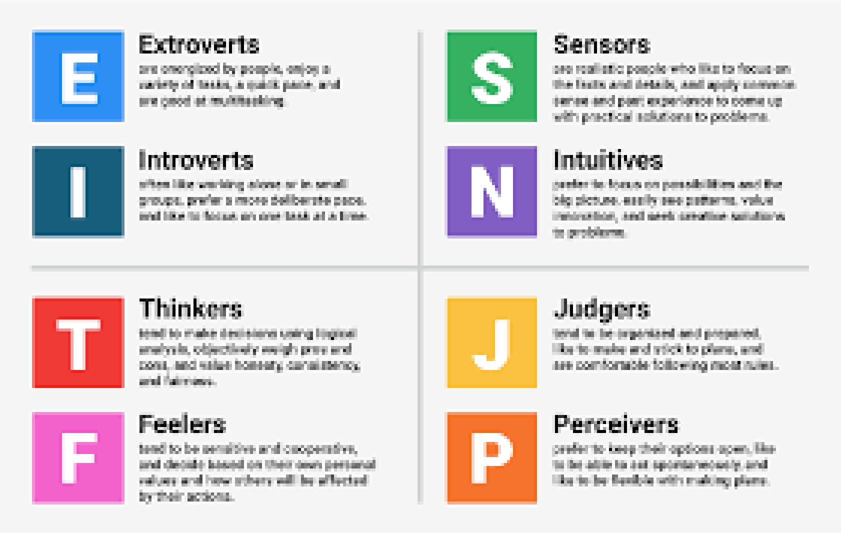
Social media has become the most widely used communication and interaction tool between people over the past few years. Direct interaction between people is decreasing as people tend to communicate indirectly through smartphones. Thus, it is quite difficult to recognize person’s personality. However, what’s written in social media might help us to get the information needed as people spend much time checking social media and expressing their feelings and thoughts through statuses, comments, and updates.

The project is based on identifying the personality of an individual using machine learning algorithms and big 5 models. The personality of a human plays a major role in his personal and professional life. Nowadays, many organizations have also started shortlisting the candidates based on their personality as this increase the efficiency of the work because the person is working in what he is good at than what he is forced to do.

The **Big Five model** is also known as the **Five-Factor Model (FFM**) and **OCEAN model** was developed in the early 1980s according to many psychological theories. When the statistical analysis is applied to personality survey data, some words used to describe the person and these words give a summary of the overall character or personality of the person accurately.

- **Open to Experience:** It involves various dimensions, like imagination, sensitivity, attentiveness, preference to variety, and curiosity.
- **Conscientiousness**: This trait is used to describe the carefulness and diligence of the person. It is the quality that describes how organized and efficient a person is.
- **Extraversion:** It is the trait that describes how the best candidates can interact with people that is how good are his/her social skills.

- **Agreeableness:** It is a quality that analyses the individual behaviour based on the generosity, sympathy, cooperativeness and ability to adjust with people.
- **Neuroticism:** This trait usually describes a person to have mood swings and has extreme expressive power.



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.

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.

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.

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.

Personality Traits(Fig.)

TECHNOLOGY USED-

MACHINE LEARNING

Machine learning (ML) is a type of artificial intelligence that allows software applications to become more accurate at predicting outcomes without being explicitly programmed to do so. Machine learning algorithms use historical data as input to predict new output values. ML is one of the most exciting technologies that one would have ever come across. As it is evident from the name, it gives the computer that makes it more similar to humans: ***The ability to learn***. Machine learning is actively being used today, perhaps in many more places than one would expect.

Research on personality type prediction from textual data is scarce. However, important steps have been taken in this endeavour through machine learning. Classic machine learning techniques and neural networks have been used successfully for predicting MBTI personality types.

- Machine learning uses data to detect various patterns in a given dataset.
- It can learn from past data and improve automatically.
- It is a data-driven technology.
- Machine learning is much similar to data mining as it also deals with the huge amount of the data.

FIELD OF PROJECT-

Simple demonstration of using data mining concepts and python data science libraries to analyse and classify personalities of a given set of people or an individual. Apart from this, Personality Psychology, demographics are used to showcase the prediction of personality depending of factors for which it is used.

TECHNICAL TERMS USED –

In our project, we have tried to combine both personality prediction using ML algorithms **like KNN, CNN and Logistic regression** to predict the personality of a person and used term frequency algorithm to get the skill in which the person is good at. Users can easily identify his personality and his technical skill from this model or system.

We have treated personality trait detection as a **supervised learning problem**. Our process was to train four classifiers to predict a binary outcome for each of the four **MBTI types**. So essentially for an excerpt of text, each classifier would predict one of the four dimensions for the MBTI personality types.

Classification, Regression, Cluster analysis, Association.

LITERATURE SURVEY-

1. Maite et. al. [7] focused on Personality prediction from the Author Profiling task. They used PAN-AP-2015 corpus that was collected for social media users from twitter. Four languages were included but this paper focused on English language only. Self- online test was taken, and score was given between -0.5 to 0.5. Big Five model was used for traits. Then Glove representation in vector form was used for word embedding. For short

   input data, the padding of many zero was done to as CNN require fixed amount of input. Different filters were used for Convolution layers. All the outputs were merged together, and the pooling layer was applied. Re LU is used as activation function. Fully connected neural network gives output as 5 neurons one for each stage. Deeper CNN can be implemented.

   The authors in this paper [8] aim to predict the personality of twitter users for Arabic users in Egypt. They collected the data using Ara Personality. This data set was collected from Arabic dialect twitter user. Questionnaire consisting of several MCQs having 5 choices were translated to Arabic language and then filled by the users. And scores were assigned to each choice chosen by the user on the basis whether the question is Proportional or inversely proportional to the Big Five Personality Traits abbreviated as OCEAN. Apart from questionnaire their feeds were also collected. These Collected users feed then were pre-processed and cleaned by removing noisy data like user names, emails etc. and some non-Arabic words were converted to Arabic. Normalization was done to keep all the words in one form. The data is then divided into Train and test data. TF-IDF was calculated for every user. Three

Supervised Machine learning as algorithms namely Decision trees, Support vector Machine and Multinomial Nave Bayes was used.

M. Hassanein et. al. [9] presented an approach to predict the personality on basis of semantics. They used big five model on My Personality Data-set. Vector Space model is used to represent the user text in the vector from that hold counts of every word in the text. Similarity measure is used to measure semantics using WordNet Database.

The Authors of the paper [10] proposed the model for text analysis and predict the personality of brands on Social Media Platform. Big Five model was used to predict the brand personality. This information could help brand to plan its Marketing Strategies as well as Improve relations with the Customers. My Personality data-set was used as well as the one was created for Brands pages and features were extracted from both these data-sets. Feature selection was done by done approaches namely Pearson Correlation and other was Gradient Boosting on 3 different Machine learning approaches as Support Vector Regression (SVR), Gradient Boosting and Feed-Forward Neural Network. XGB models perform best and predict personality.

2. CLSTM that is a bidirectional Long Short-Term Memory network interconnected with CNN to find personality of users.

   It focused on structure of text as it can be important feature. Big Five model with 5 traits was used. Two data- sets were used for the experiment. One is long text data- set of essay data-set of 2467 essays tagged with their authors traits and another is Short text of YouTube vloggers. GloVe algorithm was used for word embedding. LSTM is used which has a self-loop and RNN loop as well, it is bidirectional so as to extract more features.

   Paper also proposes the concept of Latent sentence Groups (LSG) that means several sentences that are closely related to each other. CNN was used for studying such latent features. Max pooling layer is used after LSTM to get sentence vectors. Soft max classifier was used as the classifier. Various contrast models were used like TF-IDF bayes, 2 and 3 dimensions CNN, one LSTM to compare the results with proposed model, which proved to perform better.

   The authors of this paper [12] presented a system that could analyze the personality traits for Facebook users by using their status posts. Big Five personality model was used. They used My Personality data-set that had 250 users and about 10,000 posts updates from these users. These posts after extraction were pre-processed by removing links, symbols etc. All the words were converted in their lowercase. A spelling correction algorithm was used for real time data to correct all the incorrect spellings in the post. Posts also consisted of symbols like Hashtags (#) and emotions, these were removed by keeping the words as it is. TF-IDF was calculated to extract keywords from documents, thus feature vector was formed. This vector was too large so to reduce the size and to get only relevant features, Principal Component Analysis was used.

   Machine learning algorithms KNN and SVM were used. KNN was best for Classification of traits.

Previous study on personality prediction has been done by using social media Facebook and some features such as LIWC features, SNA features, time-related features. Their research is very similar with ours especially for the dataset (250 dataset from my Personality) and the features (LIWC and SNA features). Another research in personality prediction based on Facebook status were done by using two approaches such as open vocabulary DLA (Differential Language Analysis) and LIWC features. By using Facebook, a research defining features with bag-of-words and token (unigrams) approaches were conducted as well. Other study was done to make a personality prediction system by using Twitter with LIWC and MRC as featured . All mentioned above researches did personality prediction by using social media in English based on Big Five Personality models. Recent research was conducted to make a personality prediction system using Twitter in Bahasa based on Big Five Personality models. Other research on personality prediction was done using deep learning technique to classify Big Five Personality models from social media .

METHODOLOGY-

1) GATHERING DATA-

   This step includes the below tasks: Identify various data sources, Collect data. Integrate the data obtained from different sources. By performing the above task, we get a coherent set of data, also called as a dataset.

2) DATA PREPARATION-

   This step can be further divided into two processes:
   Data exploration:
   It is used to understand the nature of data that we have to work with. We need to understand the characteristics, format, and quality of data.
   A better understanding of data leads to an effective outcome. In this, we find Correlations, general trends, and outliers.
   Data pre-processing:
   Now the next step is preprocessing of data for its analysis.

3) DATA WRANGLING-

   It is not necessary that data we have collected is always of our use as some of the data may not be useful. In real-world applications, collected data may have various issues, including: Missing Values, Duplicate data, Invalid data, Noise.

4) DATA ANALYSIS-

   Now the cleaned and prepared data is passed on to the analysis step. This step involves:
   Selection of analytical techniques
   Building models
   Review the result

5) Train Model

Now the next step is to train the model, in this step we train our model to improve its performance for better outcome of the problem. We use datasets to train the model used in various machine learning algorithms. Training a model is required so that it can understand the various patterns, rules, and, features.

6) Test Model

Once our machine learning model has been trained on a given dataset, then we test the model. In this step, we check for the accuracy of our model by providing a test dataset to it. Testing the model determines the percentage accuracy of the model as per the requirement of project or problem.

7) Deployment

The last step of machine learning life cycle is deployment, where we deploy the model in the real-world system. If the above-prepared model is producing an accurate result as per our requirement with acceptable speed, then we deploy the model in the real system. But before deploying the project, we will check whether it is improving its performance using available data or not. The deployment phase is similar to making the final report for a project.

FACILITES REQUIRED FOR PROPOSED WORK-

HARDWARE AND SOFTWARE USED –

- **Operating system that you prefer (it could be Linux, Mac or Windows)**
- A PC or laptop that you own
- Your PC or laptop should have enough RAM to run the calculations uninterruptedly.
- Python 3 or above installed in your system.
- Some basic machine learning algorithms

CONCLUSION-

This paper provided an insight on existing attempts of the task of personality prediction from text on social media to-date, along with the various kinds of social medias which have been utilized for said task. While most personality prediction studies to-date require a dataset to perform supervised learning, it is costly to obtain a dataset labelled with personality traits of social media users. Recent studies have tried applying semi-supervised and unsupervised learning to tackle this problem. Further improvements to the existing state of personality prediction can be made by expanding the target language, applying more suitable algorithms or preprocessing methods to achieve higher accuracy, and implementing said task to other personality models that can be used to predict a person's personality with an accuracy of 85.81%**.** Used to identify the right candidate to the right candidate based on his personality and skill.

Behaviour on Social media sites of users can help in predicting the traits of User based on various personality models. Earlier questionnaire method was used that could be a Costly and time-consuming process. The goal of this paper is to give summary of the work done for Predicting the personality on text from Social media sites and to summarize the future trends. Table I Shows the Overview of the Current research techniques Performed analysis shows the Various techniques and models used. Working on the future directions, accuracy can be increased of prediction as well as can be used to provide some Customized services and other recommendations**.**

REFERENCES-

- Diener, E. and Lucas, R. (2019). Personality Traits. [online] Noba. Avail- able at: https://nobaproject.com/modules/personality-traits [Accessed 30 Sep. 2019
- https://www.geeksforgeeks.org/overview-of-personality-prediction-project-using-ml/
- https://www.kaggle.com/c/twitter-personality-prediction/overview
- https://www.sas.com/en_in/insights/analytics/machine-learning.html#:~:text=Machine%20learning%20is%20a%20method,decisions%20with%20minimal%20human
- https://www.researchgate.net/publication/316254176_Exploring_personality_prediction_from_text_on_social_media_A_literature_review
- https://www.javatpoint.com/machine-learning
- https://www.ijert.org/personality-prediction-from-social-media-text-an-overview