INLS625: INFORMATION ANALYTICS

PROJECT PHASE-I CONCEPT REPORT

VIBHOR GUPTA

2022-09-29

Proposal – 1:

What’s your Personality? Let’s find out using Information Analytics.

Problem Definition:

Personality is the combination of unique characteristics that defines a person, and predicting other people’s personalities is always a favorite thing for humans. This benefits us in maintaining a positive relationship with others. Think of your friend circle; you are not the same with everyone. There will be minute differences in your behavior around each of them, and this helps you to maintain a healthy friendship with them. In this technology-driven era, all the big tech firms, from social media, or search engines, or retail stores, are also trying to predict the personality of the users. By getting a sense of user persona, they want to capture the user. E.g., Take the case of big retail giant Walmart, which tries to capture the user’s persona by looking at their buying habits and then will recommend products and deals based on those. This helps users to become loyal to the giant and increase their business.

My goal for this course project is to look at how to capture user personas from the data they generate. For this purpose, I came upon two datasets I will explain in the Data Understanding section. I will be using one of them for this project. Another motivation to push towards this project is curiosity; I got introduced to a few alien terms like INFJ, INTP, etc., and using Google, I found out they represent people's personalities. I got interested in looking more into the topic and thought it would be a good project for this course.

Data Understanding:

I came upon two separate datasets for personality prediction. One is the Big 5 personality data from the openpsychometrics website, which captures user persona into five traits abbreviated as OCEAN. The dataset asks questions from users and tries to cluster them. This will be an unsupervised type of approach.

The other dataset is the Myers–Briggs Type Indicator, which captures users' personalities into one of 16 types. The dataset captures users' comments and tries to predict their personalities based on them**.**

# Both datasets have a decent size, as mentioned in the Data Preparation module below.

Data Preparation:

The first dataset has around 1015341 records, with 110 columns. No target is available for prediction.

The second dataset has around 8675 records, with a personalized label and the comments for that personality. This will mostly come in the domain of text mining and will require NLP techniques to solve the problem.

As for data preparation, since the datasets are already available, it is required to look at the distribution of attributes for different personalities.

For both datasets, it is required to convert the data into a machine-interpretable format and then feed it into the algorithms.

Modeling:

Both datasets require different sets of models. Unsupervised tasks require models to cluster the data points like KNN, K-Means, hierarchal clustering, etc. For the supervised job, I will focus on decision trees, SVM, logistic regression, etc.

As for the tools are concerned, I will be focusing on using WEKA/Orange for quick data analysis, but since I am a little bit proficient with python, I will be using it for a significant part of the project.

Evaluation and Deployment:

I am not trying to reinvent any wheel here. The datasets are widely studied and have publications associated with them. For evaluation purposes, I can compare different models I will create or my models’ performance with the available publications.

As far as the model deployments are concerned, this is a long goal, but I plan to develop a web app that users can look at and maybe interact with to learn about the personalities.

The primary learning from the project is to utilize information analytics to come up with personality predictions of users. The future goal is to build on this project to use digital footprints for personality prediction.

Risk Mitigation:

I am not committing to any accuracy or performance measure for the model. There are a few risks associated with the model's performance that it will throw up random predictions in a bizarre manner. Apart from this, I can’t think of anything else.

References:

* Big 5 data can be found on the [link](https://www.kaggle.com/datasets/tunguz/big-five-personality-test)
* MBTI data can be found on the [link](https://www.kaggle.com/datasets/datasnaek/mbti-type)
* MBTI [foundation](https://www.myersbriggs.org/my-mbti-personality-type/mbti-basics/)
* Personality Type Based on the Myers-Briggs Type Indicator [link](https://doi.org/10.48550/arXiv.2201.08717).

NOTE:

* I am also new to predicting persona using user-generated data, so I have limited information in the area. But I hope to expand my knowledge as we progress with the project.
* I will choose one of the datasets after reading more about them, or I may be able to find a new one.
* I am aiming for two project proposals for the course. The reason for giving two suggestions is that I am working independently with Prof. Javed on a text mining project. The text mining task is long enough, but we plan to complete it this semester. That will continue throughout the semester, and if something tangible comes out of this, I am willing to submit it as a course project ( already discussed with Prof. Javed ).
* The personality prediction idea is a fallback, which I will work on. This will also be an independent project, which I will be pursuing to test my skills and learn something new.