Personality Prediction System from the textual data taken from multiple social media sources

Iram Nawab
iramnawab@iitgn.ac.in

Shoaib Alam
shoaibalam@iitgn.ac.in

Darshan Patil
patildarshan@iitgn.ac.in

Venkata Sai Kumar Gavini
saikrgavini@iitgn.ac.in

Shriraj Sawant
sawant_shriraj@iitgn.ac.in

1 INTRODUCTION

In recent years, massive amount of data has been generated by the users via social media especially in the form of textual data type. People often update statuses, posts, comments to express their feelings and opinions on social media. These expressions can be used to characterize the individual's behavior, personality, and characteristics of their thought patterns. In fact, earlier studies in psychology show that there is a strong correlation between user personalities and their linguistic behavior. Personality prediction has important practical applications in diverse areas ranging from recruitment systems, personal counseling systems,computational advertising, marketing science, enhanced human-computer interaction, and bank credit scoring systems.

In this work, we intent to develop a system that can predict one's personality based on the post or status one uploads on the social media network. This study is based on the prediction of Big Five Personality, MBTI (Myers-Briggs Type Indicator), i.e., I-E: Extraversion (E) or Introversion (I), N-S: Sensing (S) or INtuition (N), T-F: Thinking (T) or Feeling (F), J-P: Judging (J) or Perceiving (P). To date, there are many models which try to predict human personality; they use the older machine learning algorithms such as RNN and LSTM. These models take very high time to train due to sequential inputs, and also, they are not good enough to capture the semantic meaning of the words, which is the loss of context. We are trying to develop a model which can outperform the existing models using the pre-trained language models.

2 RELATED WORK

Personality prediction using social media is not new. For example, research conducted by [1], use dataset obtained from MyPersonality that consists of data of 250 Facebook users with around 10,000 statuses and maps to personality label based on the Big Five Personality Traits model. The second dataset is collected manually using Facebook API Graph that consists of statuses of 150 Facebook users. Linguistic Inquiry and Word Count (LIWC) and Structured Programming for Linguistic Cue Extraction (SPLICE) are used as Linguistic feature extraction methods. Social Network Analysis (SNA) provided by myPersonality dataset in form of detail information about social structures such as user's friendship network is also used. Traditional machine learning algorithms such as Naive Bayes, Support Vector Machine (SVM), Logistic Regression, Gradient Boosting, and Linear Discriminant Analysis (LDA) and deep learning implementations using four architectures, MLP (Multi-Layer Perceptron), LSTM (Long Short Term Memory), GRU (Gated Recurrent Unit), and CNN 1D (1-Dimensional Convolutional Neural Network) are used in the classification problem. Experimental results showed while Openness (OPN) and Extraversion (EXT) has the highest average accuracy in myPersonality dataset and manual gathered dataset respectively ,there is no architecture that dominated all big 5 personality traits.

Another study [2] attempt to find optimal setting to perform personality prediction with the dataset taken from an unpublished manuscript (V. Ong, A. D. S. Rahmanto, Williem, D. Suhartono and E. W. Andangsari, "Personality Modelling of Indonesian Twitter Users with XGBoost Based on the Five Factor Model") correlation-based feature subset selection is used to minimize the dimension. To separate train and test data 10-fold cross validation is used. Three comparisons: amount of n-gram, twitter metadata, and classifier used are analyzed. The optimal result around the usage of 3000–4000 word n-grams is obtained with the highest F-average of 0.7482 . Random forest and SMO performs well with our dataset. However classifiers like KNN is suspected to be prone to outliers in the dataset. The major issue of the study is the small number of dataset used in this study.

In [3], researchers use multiple social media data sources (Facebook and Twitter) to produce a predictive model for each trait.Multi model deep learning architecture is introduced by combining the statistical based text feature and a predefined model feature to improve the performance in predicting a personality of a person.

3 DATASET

The dataset used in this study is from the Myers–Briggs personality dataset which is openly available on Kaggle. The Dataset Myers–Briggs Type Indicator (MBTI) is a psychological classification of humans that they experience through the four basic principles of psychological functions i.e sensation, intuition, feeling, and thinking.

The dataset consists of 8675 unique values in which the first column is for the MBTI personality type of a particular person, it is a four-letter MBTI code/type. The second column is a section of each of the last 50 things they have posted. These entries are separated by '|||' (3 pipe characters).

This data has been collected from the users of an online forum (PersonalityCafe forum), this forum provides a large section of people, where each person is given a questionnaire that recognizes their MBTI personality type.

4 PROPOSED SOLUTION

The working procedure of this proposed system can be classified into the following stages.

(1) **Collection of dataset and their re-sampling:** We will be collecting data from Kaggle, which has 8675 rows, and each row is of a unique user. Each row user's last 50 social

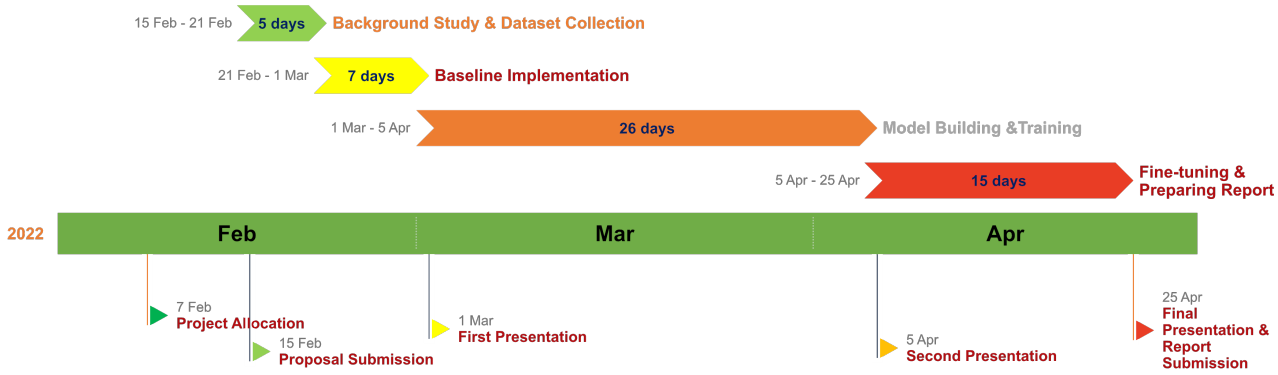

Figure 1: Proposed Project Timeline

media posts and their MBTI personality type. So, as a result, we will have 422845 records obtained from this dataset.

(2) **Re-sampling:** The original dataset is skewed, and the data are not evenly spread across all four dichotomies, I/E Trait, S/N Trait, T/F Trait, and J/P Trait. To remove this imbalance, we will do a re-sampling of data.

(3) **Preprocessing:** To get more exploration of the personality from the text, we will preprocess the data by tokenization, removal of user mentions, hashtags, stop words, URLs, and word stemming.

(4) **Text-based Classification:** This proposed solution will use a supervised learning approach. It will use the text as input and predict the personality trait (I-E, N-S, T-F, J-P) accordingly.

(5) We will use the pre-trained model BERT that has been trained using a large data corpus that includes 2.5 billion words from the Wikipedia site, and it has a dictionary that contains 800 million words. BERT architecture consists of 12 encoders layers and 768 hidden units, and 12 attention heads.

(6) **Comparing Efficiency:** We will be comparing the performance of our model with other classifiers.

(7) **Evaluation Metrics:** We will be using accuracy, f-score, recall, and precision for evaluating the model's performance.

5 PROJECT MANAGEMENT

(1) Looking for Proportionality in Dataset: In this phase we will plot the dataset to determine the distribution of the MBTI personality types in the dataset.

(2) Pre-processing and Categorizing dataset: We will remove URLs and stop words from the dataset. To categorize dataset, suppose if first category is Introversion (I)/Extroversion (E), the second category is Intuition (N)/Sensing (S), the third is Thinking (T)/Feeling (F) and the fourth category is Judging (J)/Perceiving (P). As a result, for each category, one letter will return and at the end there will be four letters that represent one of the 16 personality types in the MBTI. For instance, if the first category is returning I, the second category is returning N, the third category is returning T and the fourth category is returning J, the relevant personality type would be INTJ.

(3) Model Building Phase: During these phases we will build model on training data and predictions will made for the testing data. We will also evaluate performance of model on testing dataset.

(4) Fine tuning and creating report phase: We will compare the model with other existing methods which used the same dataset and try to increase accuracy of model.

The expected timeline of the project is given in the figure 1

REFERENCES

[1] Tommy Tandera, Hendro, Derwin Suhartono, Rini Wongso, and Yen Lina Prasetio, "Personality Prediction System from Facebook Users". Available at https://www.sciencedirect.com/science/article/pii/S1877050917320537, 2017

[2] Nicholaus Hendrik Jeremy, Cristian Prasetyo, and Derwin Suhartono, "Identifying Personality Traits for Indonesian User from Twitter Dataset", Available at https://doi.org/10.5391/IJFIS.2019.19.4.283, 2019

[3] Hans Christian, Derwin Suhartono , Andry Chowanda and Kamal Z. Zamli, "Text based personality prediction from multiple social media data sources using pre-trained language model and model averaging",Available at https://journalofbigdata.springeropen.com/articles/10.1186/s40537-021-00459-1, 2021