Psychometric Tester and Analyzer

M.D. Sale[1], Atharva Sahu[2], Rasika V Burde[3], Prachi[4]

[1]Asst. Professor, Dept of Computer Engineering, Sinhgad College of Engineering, Vadgaon, Pune-411041, Maharashtra, India

[2, 3, 4]B.E. Student, Dept of Computer Engineering, Sinhgad College of Engineering, Vadgaon, Pune-411041, Maharashtra, India

Abstract

Personality is the characteristic patterns of thoughts, feelings, and behaviors that make a person unique. Our persona substantially impacts our lives. Nowadays, predicting personality from social media is transforming into a trending research area in computational linguistic. Traditionally, personality assessment has always been performed by psychology experts, with the help of interviews or self-reports. Nowadays, with the help of social media platforms, more and more people can express their activities, feeling, thoughts, and opinions. Posts, comments and status updates made by users of social media can reveal personal information. Researchers have exploited this data to reveal patterns, thought-process, behavior, preferences, persona, and other informative values. Twitter, being the most popular social media used currently, is used here for speculating and analyzing data. The research can be further advanced by using different techniques to classify and unveil more statistics/data. Here we are using machine learning techniques, such as NN and NLP, to predict personality traits.

Keywords

Social Media, Personality Traits, Machine Learning, LSTM, BERT base, NLP,NN

I. PROPOSED WORK

Social media is booming, with millions of people posting messages, tweets or photos every minute. According to a recent survey, Twitter has only 353 million monthly active users worldwide, and 9,281 tweets and 500 million tweets are sent in one second. Users reveal a lot about themselves by creating their profile on social networks, they showcase their personalities through posts, status, updates, interests which can be extracted from their profile. Personality is a typical pattern of thoughts, emotions, and behaviors that make a difference. Personality affects decision-making and human behavior. Traditionally, personality assessment is usually conducted through interviews by experienced psychologists. But nowadays, technological advancements provide us with a platform to bridge the gap between personality research and social media. When examining previous work in this field, we found that the information in Facebook user profiles is not an "idealized" version of their own, but reflects their true personality [1]. Similarly, we assume that other users of social media platforms are doing the same thing, and Twitter (with 188 million users online every day) becomes an ideal research platform.

Going through Previous works done in this field, we saw that the information in users' Facebook profiles is not an “idealized” version of themselves, but a reflection of their actual personalities [1]. Similarly, we assume that other users of social media platforms will do the same, and Twitter is one of the most popular daily users, with 188 million users, making it an ideal research platform. Developers to allowed access twitter's data. Python library TWEEPY can be used to create an interface between twitter and python, allowing access to twitter's data. Our research focuses on social behavior and language habits on Twitter. By extracting textual data, then processing and classifying them we can gain insights on user's personality. We proposed a systematic system, which is based on two models: first, LSTM, which is based on recurrent neural network (RNN). Second, BERT base uncased, which is pre trained neural network-based technique for natural language processing. We checked the predictive nature of predictive features using personality characteristics. We have explored a lot of information that does not require further processing, so the accuracy and efficiency of prediction may be higher. We have surveyed and selected the best algorithms for higher predictions.

Our research has made some contributions to predicting accurate data. Firstly, to study connections between user' information on social media and their actual personality [16]. Second, to find out the traits of social networks users using MBTI personality types. Third, providing data to demonstrate that pre-trained BERT base model provides higher accuracy than other models.

II. LITERATURE SURVEY

The article proposes a personality prediction system based on neural network and natural language processing. The proposed system can use user data and more accurately reveal the user's persona. People use social media to freely talk about topics related to their life and family happiness, psychology, their interaction with society and the environment, and politics [2],[4]. Previous persona evaluation strategies that used RNN and LSTM is juxtaposed with pre-trained language model BERT (Bidirectional Encoder Representations from Transformers) [3].

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. It explains a new language representation model that can be customized with an additional output layer to create a model that is superior to other natural language processing models. [5].

Refining Word Embeddings Using Intensity Scores for Sentiment Analysis. IEEE. It introduced a new method that uses intensity estimates from sentiment lexicons to improve word attachments for sentiment analysis. In “A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis”, effectiveness of combining emoji preprocessing and user-generated content is emphasized. Due to the many factors involved, the classification of personality is difficult, and even a person can classify a person inaccurately based on the text. However, the previously trained language model can capture the different subtleties of language use of different personality types. [6].

A Framework for Generating User Embedding. Explains sentence representation using BERT. Their encoding was used for two classification tasks: depression detection and personality classification based on Reddit dataset [11].

Table 1- Comparison Table

Paper Title	Year	Seed Idea
Facebook Profiles Reflect Actual Personality, Not Self-Idealization	2010	-People express their original thoughts on social media instead of an idealized version of themselves
Personality traits recognition on social network—Facebook	2013	-Correlation between user's personality and social media is strong
The development and psychometric properties of LIWC2015	2015	-the work of personality extraction from the text.
Neural Networks in Predicting Myers Brigg Personality Type From Writing Style	2017	-LSTM with other model gives more accuracy. LSTM has potential of More information to be encoded in parameters
Refining Word Embeddings Using Intensity Scores for Sentiment Analysis. IEEE	2018	-They introduced a new method that uses intensity estimates from sentiment lexicons to improve word attachments for sentiment analysis
A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis	2018	-Emphasize on the effectiveness of the combination of emoji preprocessing and user-generated content -, lemmatization was also shown to have beneficial effects on accuracy
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding	2019	-Natural language processing with BERT obtains state-of-the-art results.
Myers-Briggs Personality Classification and Personality-Specific Language Generation Using Pre-trained Language Models	2019	-Larger and cleaner data sets will increase the accuracy of BERT -BERT model for prediction provide higher accuracy than other models.
The impact of preprocessing steps on the accuracy of machine learning algorithms in sentiment analysis.	2019	-The influence of preprocessing on the accuracy of the three machine learning algorithms, namely NB, SVM and Maximum Entropy (MaxE), was checked, and it was shown that the accuracy was significantly improved in the case of NB. -The accuracy after preprocessing the text in the application phase improved significantly
Personality Prediction using Twitter Data	2020	-Predicting the personality of an individual based on BERT model using Twitter. -Used Xg-boost algorithm to compute results.
Author2Vec: A Framework for Generating User Embedding	2020	-Sentence representation using BERT -Their encoding was used for two classification tasks: depression detection and personality classification based on Reddit dataset.
Cross-Domain Sentiment Encoding through Stochastic Word Embedding. IEEE	2020	-The new technique of random nesting is applied to classify the sentiment between domains while maintaining the similarity of the embedding.
Sentiment analysis using deep learning architectures: A review.	2020	-CNN does a better job in computer vision and image processing, and the same architecture is often used in word processing
Text based personality prediction from multiple social media data sources using pre-trained language model and model averaging.	2021	-Deep learning approach with BERT outperforms most personality model -Addition of other NLP statistical feature with BERT can increase model performance.

III. PROPOSED WORK

In this research, firstly, we extracted our datasets from Kaggle Twitter, social networking platform, and then datasets are sent for preprocessing. Further, the preprocessed data is sent to LSTM and BERT-base Uncased models for classification, resulting in personality types. The personality types are predicted with higher accuracy and reasons for higher accuracy will be unveiled further.

3.1 Dataset Description

The static dataset used here is the famous **Myers-Briggs Personality** Type Dataset that includes a large number of people's MBTI type and content written by them. This dataset contains over **8600** rows of data, on each row is a person's:

- Type (This persons 4 letter MBTI code/type)
- a section of each of the last 50 things they have posted (Each entry separated by "|||" (3 pipe characters)).

The dynamic dataset is extracted from twitter using twitter API keys, which allow us to obtain twitter data with the permission of Twitter's owner. This extraction is commonly known as web scrapping. We are going to extract 50 recent (texts only) tweets of users. TWEEPY, i.e., python library for connecting with Twitter API. Web scrapping provides us with dynamic data that is further sent for processing.

3.2 Data Preprocessing

The dataset is preprocessed before using it for any analysis. This is done in order to make system more efficient and accurate. We have used regex to detect special characters like '@, emojis' etc. from the posts, remove stop words and punctuation, convert the text to lowercase and stemming to extract the root of words. The pre-processed data is split using train_test split and sent to the Keras model (BERT and LSTM) for predictions.

3.3 Data Analysis and Predictions

After pre-processing, we split our cleaned dataset into three subsets i.e., training, validation, and testing set, using sklearn model selection. Model selection is basically a way to set a blueprint to analyse data and then using this data to measure new data for classification and model selection.

3.3.1 LSTM

In our research, we have used LSTM as a baseline model. LSTM is based on neural network, basically recurrent neural network. LSTM is capable of handling long-term dependencies; It runs in loops and error gradient decreases. Hence, we used LSTM. LSTM also memorises the content in a good way and this memorised content is used for future predictions.

Now we have used keras layer's API to import all the libraries of NN in keras. These layers are building block of neural network and they consist of a tensor-in, tensor-out computation function. Then, we vectorized our data, by turning text into a sequence of integers/vectors. Further, we use text_to_sequences to find the most frequent words in texts, known by the tokenizer, and replace it with corresponding integer value from word_index dictionary. After that, we perform one-hot encoding to eliminate hierarchies' issues. Now, we use model.fit() from sklearn, which takes the training data as arguments and returns predicted labels. The function model.fit() uses a hyper parameter epoch, which defines number of times the learning algorithm works through the entire training set. As the number of epoch increases, accuracy of our model also increases.

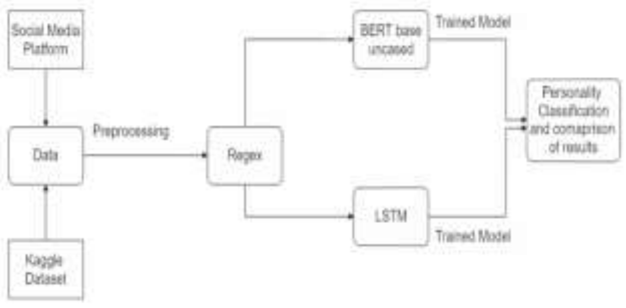
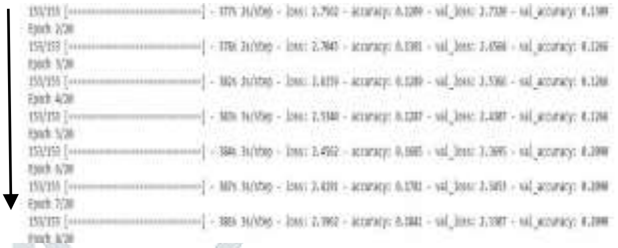


Fig. 1 Proposed System Overview Diagram



```
153/153 [==============================] - 177s 1s/step - loss: 2.7502 - accuracy: 0.1200 - val_loss: 2.7330 - val_accuracy: 0.1500
Epoch 2/20
153/153 [==============================] - 178s 1s/step - loss: 2.7047 - accuracy: 0.1301 - val_loss: 2.6500 - val_accuracy: 0.1200
Epoch 3/20
153/153 [==============================] - 182s 1s/step - loss: 2.6159 - accuracy: 0.1200 - val_loss: 2.5360 - val_accuracy: 0.1200
Epoch 4/20
153/153 [==============================] - 180s 1s/step - loss: 2.5140 - accuracy: 0.1287 - val_loss: 2.4387 - val_accuracy: 0.1200
Epoch 5/20
153/153 [==============================] - 184s 1s/step - loss: 2.4562 - accuracy: 0.1685 - val_loss: 2.3695 - val_accuracy: 0.2000
Epoch 6/20
153/153 [==============================] - 187s 1s/step - loss: 2.4201 - accuracy: 0.1781 - val_loss: 2.3453 - val_accuracy: 0.2000
Epoch 7/20
153/153 [==============================] - 188s 1s/step - loss: 2.3962 - accuracy: 0.1841 - val_loss: 2.3387 - val_accuracy: 0.2000
Epoch 8/20
```

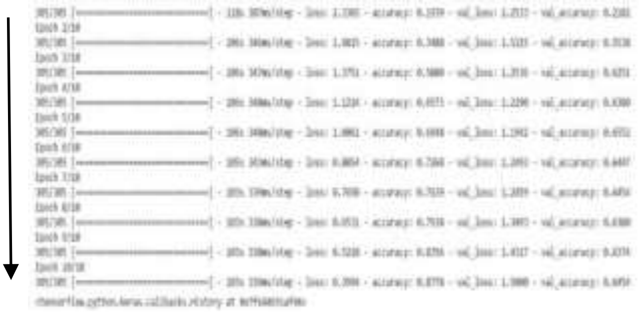
Fig.2 Increment of accuracy with epoch (LSTM)

3.3.2 BERT base

We emphasize here on BERT as we have used it to increase accuracy of our system. BERT computes vector representations of natural language that are suitable for our deep learning model. BERT is basically a transformer-based machine learning technique for natural language processing developed and pre-trained by google.

The processed and clean data is passed through pad_sequences to make it of optimal length. As a result, we receive vectorized data. Then, this data is passed through autoclasses- autotokenizer.from pretrained() to import BERT base model. This automatically retrieves relevant model, the given name to the pretrained weights.

Like with LSTM model, the data is then sent to model.fit of the keras library and again with increase in epoch the accuracy increases but this time the accuracy increases significantly.



```
305/305 [==============================] - 118s 387ms/step - loss: 1.7305 - accuracy: 0.3779 - val_loss: 1.2512 - val_accuracy: 0.2381
Epoch 2/10
305/305 [==============================] - 106s 346ms/step - loss: 1.4025 - accuracy: 0.3488 - val_loss: 1.5335 - val_accuracy: 0.3538
Epoch 3/10
305/305 [==============================] - 106s 347ms/step - loss: 1.3751 - accuracy: 0.5000 - val_loss: 1.2518 - val_accuracy: 0.6251
Epoch 4/10
305/305 [==============================] - 106s 348ms/step - loss: 1.1234 - accuracy: 0.6571 - val_loss: 1.2298 - val_accuracy: 0.6300
Epoch 5/10
305/305 [==============================] - 106s 348ms/step - loss: 1.0061 - accuracy: 0.6888 - val_loss: 1.1961 - val_accuracy: 0.6552
Epoch 6/10
305/305 [==============================] - 105s 343ms/step - loss: 0.8654 - accuracy: 0.7268 - val_loss: 1.2051 - val_accuracy: 0.6447
Epoch 7/10
305/305 [==============================] - 103s 339ms/step - loss: 0.7038 - accuracy: 0.7539 - val_loss: 1.2059 - val_accuracy: 0.6456
Epoch 8/10
305/305 [==============================] - 103s 338ms/step - loss: 0.6531 - accuracy: 0.7838 - val_loss: 1.3093 - val_accuracy: 0.6380
Epoch 9/10
305/305 [==============================] - 103s 338ms/step - loss: 0.5238 - accuracy: 0.8256 - val_loss: 1.4117 - val_accuracy: 0.6374
Epoch 10/10
305/305 [==============================] - 103s 338ms/step - loss: 0.3984 - accuracy: 0.8778 - val_loss: 1.5000 - val_accuracy: 0.6454
<tensorflow.python.keras.callbacks.History at 0x7f5480[illegible]>
```

Fig.3 increment of accuracy with epoch (BERT base)

In essence, The user data extracted is passed through LSTM and BERT based models and user's personality based on MBTI is predicted and analysed.

3.4 Personality Classification

In order to determine the classification of the user's identity, we use the Myers-Briggs method. MBTI is the most widely used personality classification used by various departments to analyze a person's personality. MBTI divides characters into four categories, namely:

- Introversion (I) or Extraversion (E)
- Intuition (N) or Sensing (S)
- Feeling (F) or Thinking (T)
- Perceiving (P) or Judging (J)

Table-2: Types of Personality Indicator

Type	Full-form
ISTJ	Introverted, Sensing, Thinking, Judging
ISFJ	Introverted, Sensing, Feeling, Judging
ISTP	Introverted, Sensing, Thinking, Perceiving
ISFP	Introverted, Sensing, Feeling, Perceiving
INFP	Introverted, Intuitive, Feeling, Perceiving
INFJ	Introverted, Intuitive, Feeling, Judging
INTJ	Introverted, Intuitive, Thinking, Judging
INTP	Introverted, Sensing, Thinking, Perceiving
ESTP	Extraverted, Sensing, Thinking, Perceiving
ESFP	Extraverted, Sensing, Feeling, Perceiving
ENFP	Extraverted, Intuitive, Feeling, Perceiving
ENTP	Extraverted, Intuitive, Thinking, Perceiving
ESTJ	Extraverted, Sensing, Thinking, Judging
ESFJ	Extraverted, Sensing, Feeling, Judging
ENTJ	Extraverted, Intuitive. Thinking, Judging
ENFJ	Extraverted, Intuitive, Feeling, Judging

Traits according to our classification:

- ISTJ traits- (Systematic, Factual, Organized, Logical, Responsible)
- ISFJ traits- (Warm, Detailed, Caring, Practical, Factual)
- ISTP traits- (Analytical, Practical, Adaptable, Curious, Problem-Solver)
- ISFP traits- (Compassionate, Aesthetic, Spontaneous, Helpful, Idealistic)
- INFP traits- (Creative, Independent, Adaptable, Inquisitive, Caring)
- INTJ traits- (Visual-oriented, Innovative, Conceptual, Logical, Determined)
- INTP traits- (Conceptual, Complex, Intellectual, Critical, Ingenious)
- INFJ traits- (Visionary, Insightful, Creative, Sensitive, Persevering)
- ESTP traits- (Adventurous, Pragmatic, Easy-going, Adaptable, Observant)
- ESFP traits- (Energetic, Sociable, Friendly, Generous, Fun-loving)
- ENFP traits- (Enthusiastic, Imaginative, Creative, Playful, Optimistic)
- ENTP traits- (Theoretical, Inventive, Abstract, Analytical, Complex)
- ESTJ traits- (Assertive, Decisive, Concrete, Active-organizer, Practical)
- ESFJ traits- (Harmonizer, Caring, Empathic, Loyal, Cooperative)
- ENFJ traits- (Appreciative, Insightful, Tactful, Imaginative, Sociable)
- ENTJ traits- (Planner, Organized, Strategist, Logical, Critical)
-

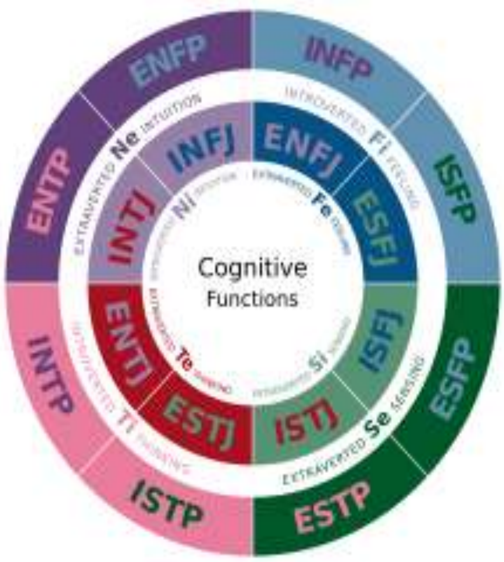


Fig.4 MBTI personality types

IV. EXPREMENTAL ANALYSIS

The project identifies individual's persona based on the posts available on their respective Twitter accounts, which can be used in various fields like criminal investigation and company recruitment processes. Our result illustrates validation accuracy of 79% and training accuracy of 85%.

Table-3: Model Accuracy

Model	Train accuracy	Validation accuracy
LSTM baseline	**18.96%**	**16.9%**
BERT-base-uncased	**85%**	**79%**

V. CONCLUSIONS

In this research, we provide an information basis for analyzing each user's social network and predicting personality. Our research is based on exploiting more information about human nature, persona, and mindsets. In order to identify the different personalities of a person, we use a variety of data sets to get an effective result. Our results show that analyzing a person's social and language networks can help gain more insights on human nature.

VI. FUTURE WORK

Although our results show improvements in other deep learning models that use language data, there are still many limitations that affect this research. The content on social media platforms is noisy. This makes the data collection, data labeling and data analysis more difficult and tedious task. The influence of using preprocessed model BERT is relatively good as accuracy increases but very large datasets when overfitted can produce incorrect results. There is a scope for generalization of this framework as the work performed here can be extended to many fields like criminal investigations and recruitment processes and also can be applied on different social media platforms.

REFERENCES

[1] M. Back, J. Stopfer, S. Vazire, S. Gaddis, S. Schmukle, B. Egloff, and S. Gosling. Facebook Profiles Reflect Actual Personality, Not Self-Idealization. Psychological Science, 21(3):372, 2010.

[2] J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn, ‘‘The development and psychometric properties of LIWC2015,’’ Tech. Rep., 2015.

[3] Christian, H., Suhartono, D., Chowanda, A. et al. Text based personality prediction from multiple social media data sources using pre-trained language model and model averaging. J Big Data 8, 68 (2021). https://doi.org/10.1186/s40537-021-00459-1

[4]Alam F, Stepanov EA, Riccardi G. Personality traits recognition on social network—Facebook. AAAI Workshop—Technical Report, WS-13-01, 2013. pp 6–9.

[5] BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding Jacob Devlin Ming-Wei Chang Kenton Lee Kristina Toutanova, Google AI Language

[6] Myers-Briggs Personality Classification and Personality-Specific Language Generation Using Pre-trained Language Models Sedrick Scott Keh, The Hong Kong University of Science and Technology

[7] Neural Networks in Predicting Myers Brigg Personality Type From Writing Style Anthony, Stanford University

[8]Refining Word Embeddings Using Intensity Scores for Sentiment Analysis. IEEEYu, L.; Wang, J.; Lai, K.R.; Zhang, X..,2018

[9]Symeon Symeonidis, Dimitrios Effrosynidis, Avi Arampatzis,A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis, Expert Systems with Applications,Volume 110, 2018,ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2018.06.022.

[10]Personality Prediction using Twitter Data, Sanjit Kumar R1, Shrivatson R G2, Rishi Priyan S3, Padmavathy T4..,2020

[11]Author2Vec: A Framework for Generating User Embedding -Xiaodong Wu* Weizhe Lin* Zhilin Wang Elena Rastorgueva University of Cambridge, United Kingdom ..,2020

[12]Cross-Domain Sentiment Encoding through Stochastic Word Embedding. IEEE Hao, Y.; Mu, T.; Hong, R.; Wang, M.; Liu, X.; Goulermas, J.Y…,2020

[13]Text based personality prediction from multiple social media data sources using pre-trained language model and model averaging Hans Christian, Derwin Suhartono, Andry Chowanda & Kamal Z. Zamli…,2021

[14]An Effective BERT-Based Pipeline for Twitter Sentiment Analysis Marco Pota , Mirko Ventura , Rosario Catelli and Massimo Esposito..,2021

[15]Sentiment analysis using deep learning architectures: A review.Yadav, A.; Vishwakarma, D.K…,2020

[16]A Survey On Personality Prediction Using Twitter M.D. Sale , Atharva Sahu , Rasika V Burde , Prachi..,2021