Predicting Human Behaviour from Social Media Activities

Abu Ekhtiar Tawhid
Department of Computer Science
& Engineering
United International University
Dhaka, Bangladesh
atawhid223018@mscse.uiu.ac.bd

Humaira Noor
Department of Computer Science
& Engineering
United International University
Dhaka, Bangladesh
hnoor222007@mscse.uiu.ac.bd

Mohammad Shafiqul Islam
Department of Computer Science
& Engineering
United International University
Dhaka, Bangladesh
mshafiquli@gmail.com

Yamir Rashid
Department of Computer Science
& Engineering
United International University
Dhaka, Bangladesh
yrashid212016@mscse.uiu.ac.bd

Md. Saddam Hossain Mukta
Department of Computer Science
& Engineering
United International University
Dhaka, Bangladesh
saddam@cse.uiu.ac.bd

***Abstract*—The purpose of this study is to investigate whether the Big Five personality traits and Schwartz values of individuals can be predicted using Linguistic Inquiry and Word Count (LIWC) variables derived from social media status updates. The researchers aim to explore the connections between LIWC factors and the target variables through Pearson correlation analysis. This analysis will allow them to examine the relationships between language patterns and personality traits/values, providing insights into the potential predictive capabilities of LIWC in identifying individuals' psychological qualities based on their online expressions.**

By collecting social media status updates as a source of linguistic data, the researchers derive LIWC variables that capture individuals' language styles and themes in their online expressions on various social media platforms. Pearson correlation analysis is employed to assess the strength and direction of the relationships between the LIWC factors and the target variables, which include the Big Five personality traits and Schwartz values.

The correlation analysis enables the researchers to identify any significant associations between specific linguistic patterns or themes derived from social media status updates and particular personality traits or values. Positive correlations would indicate that certain language patterns are related to specific traits or values, while negative correlations would suggest the absence or opposite direction of those traits or values.

Overall, this study aims to explore the potential of LIWC variables derived from social media status updates in predicting individuals' Big Five personality traits and Schwartz values. By utilizing Pearson correlation analysis, the researchers seek to gain insights into the connections between language patterns and psychological characteristics, thus contributing to our understanding of the predictive capacities of LIWC in the context of personality psychology and their relationships with personality traits/values expressed through social media.

***Index Terms*—LIWC: Big Five personality traits ;Schwartz values**

I. Introduction

This study examines the association between Big Five personality traits and Schwartz values and Linguistic Inquiry and Word Count (LIWC) variables derived from social media status updates. LIWC provides insights into language patterns and their connections to psychological processes. The goal is to assess the predictive capabilities of LIWC factors in determining individuals' values and personality characteristics.

A diverse sample population was recruited, and their social media status updates, along with self-reported measures of Big Five personality traits and Schwartz values, were collected. LIWC was applied to extract various linguistic features from the text data, including word usage, emotion, social connections, and cognitive processes. Pearson correlation analysis was conducted to examine the relationships between the LIWC variables and the target variables.

The results of the correlation analysis reveal the associations between specific LIWC variables and individuals' Big Five personality traits and Schwartz values. Positive or negative correlations indicate the direction and strength of these relationships. Significant correlations suggest the potential predictive power of certain LIWC variables for specific personality traits and values.

The findings of this study contribute to our understanding of how LIWC variables can predict individuals' personality traits and values based on their social media status updates. The Pearson correlation analysis provides insights into the specific LIWC variables that are strongly associated with the target variables.

However, it is important to acknowledge the limitations of this study, such as potential biases in self-reported data, the generalizability of the findings to larger populations, and the need for further exploration of alternative statistical techniques. Future research could employ additional validation methods to further investigate the predictive power of LIWC variables and enhance the accuracy of the models.

In conclusion, this study sheds light on the potential of LIWC variables in predicting individuals' Big Five personality traits and Schwartz values based on their social media status updates. It adds to our understanding of the relationships between language patterns, personality traits, and values in the context of social media.

II. Dataset

The dataset used in this study consists of three main components: LIWC variables extracted from Facebook status updates, Big Five personality trait scores and Schwartz values obtained from a survey.

LIWC Variables: The LIWC variables are derived from the linguistic analysis of Facebook status updates. LIWC (Linguistic Inquiry and Word Count) is a widely used text analysis tool that examines language patterns and provides insights into psychological processes. The LIWC variables capture various linguistic dimensions such as word usage, emotional expression, social connections, cognitive processes, and more.

Big Five Personality Traits: The Big Five personality traits, also known as the Five-Factor Model (FFM), are a widely accepted framework for describing human personality. The traits include Openness to Experience, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. These traits are assessed through self-reported measures obtained from survey responses.

Schwartz Values: The Schwartz values represent an influential theory of basic human values. The theory identifies ten fundamental values, including Self-Direction, Stimulation, Hedonism, Achievement, Power, Security, Conformity, Tradition, Benevolence, and Universalism. Participants in the study provide self-reported scores reflecting their endorsement of these values.

The dataset allows for exploring the relationships between LIWC variables and individuals' Big Five personality traits and Schwartz values. The aim is to determine whether specific linguistic patterns captured by LIWC can predict or provide insights into individuals' personality traits and values. The dataset provides a valuable resource for investigating the connections between language use, psychological characteristics, and individual differences in the context of social media behavior and self-report measures.

III. Methodology

A. Dataset Preprocessing

To remove outliers using the IQR method and replace them with the feature mean in a dataset containing LIWC, Big Five, and Schwartz values, you can follow these steps:

1.Identifying the subset of columns in your dataset that correspond to the LIWC, Big Five, and Schwartz values.

2.Iterating over each column and apply the IQR outlier detection and replacement process.

3.Calculating the IQR, lower bound, and upper bound for each column.

4.Replacing outliers with the mean value of the respective feature.

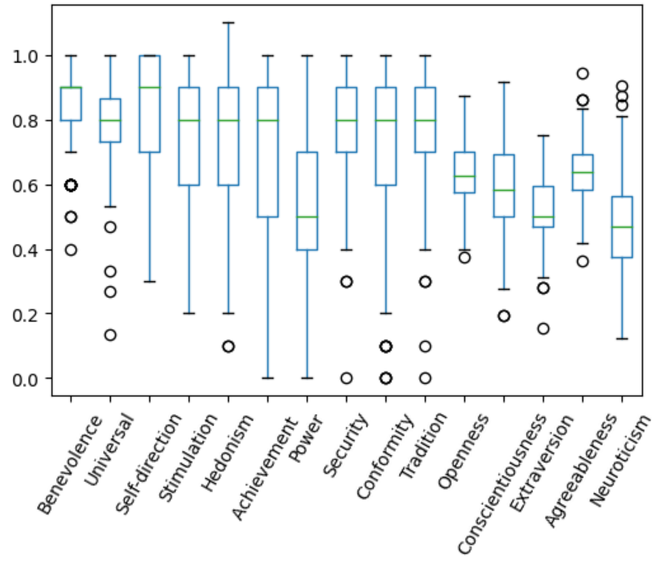


Fig. 1. Before Removing outliers

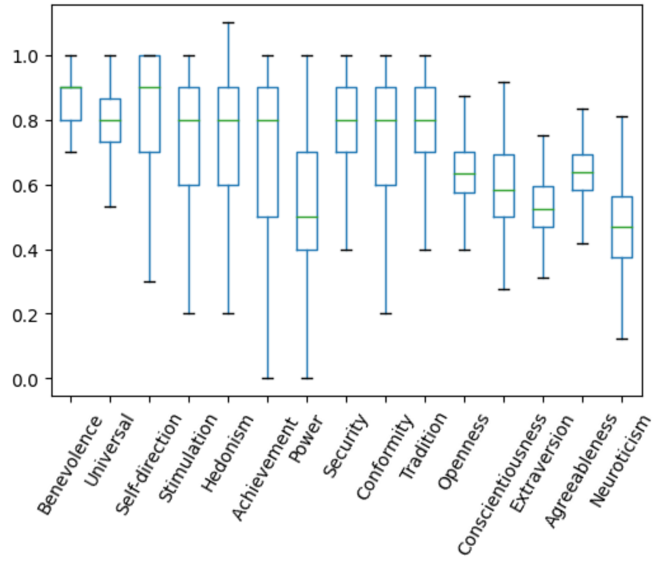
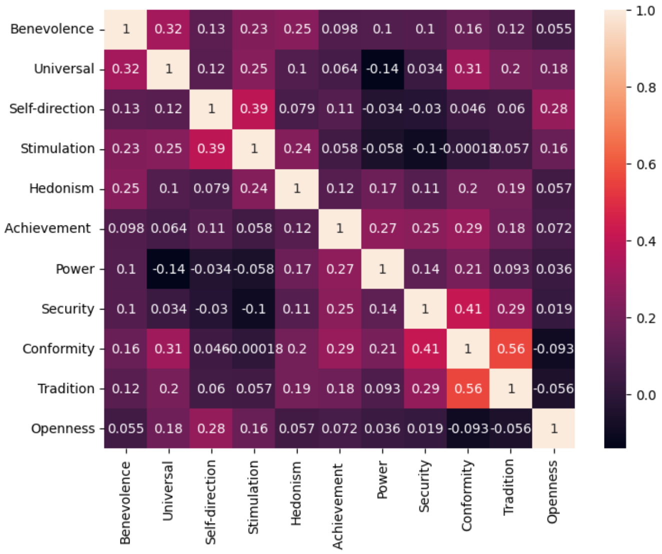


Fig. 2. After Removing outliers

B. Feature selection

Feature selection is a crucial step in data analysis to identify the most relevant features from a dataset for predictive modeling or analysis. Pearson correlation is a commonly used method for feature selection, focusing on the linear relationship between individual features and target variables.

By applying Pearson correlation for feature selection, we can assess the strength and direction of the linear relationship between each feature and the Big Five personality traits and Schwartz values. Features with a high positive or negative correlation indicate a strong linear association with the target variables, making them more likely to contribute to the prediction or analysis.



	Benevolence	Universal	Self-direction	Stimulation	Hedonism	Achievement	Power	Security	Conformity	Tradition	Openness
Benevolence	1	0.32	0.13	0.23	0.25	0.098	0.1	0.1	0.16	0.12	0.055
Universal	0.32	1	0.12	0.25	0.1	0.064	-0.14	0.034	0.31	0.2	0.18
Self-direction	0.13	0.12	1	0.39	0.079	0.11	-0.034	-0.03	0.046	0.06	0.28
Stimulation	0.23	0.25	0.39	1	0.24	0.058	-0.058	-0.1	-0.00018	0.057	0.16
Hedonism	0.25	0.1	0.079	0.24	1	0.12	0.17	0.11	0.2	0.19	0.057
Achievement	0.098	0.064	0.11	0.058	0.12	1	0.27	0.25	0.29	0.18	0.072
Power	0.1	-0.14	-0.034	-0.058	0.17	0.27	1	0.14	0.21	0.093	0.036
Security	0.1	0.034	-0.03	-0.1	0.11	0.25	0.14	1	0.41	0.29	0.019
Conformity	0.16	0.31	0.046	0.00018	0.2	0.29	0.21	0.41	1	0.56	-0.093
Tradition	0.12	0.2	0.06	0.057	0.19	0.18	0.093	0.29	0.56	1	-0.056
Openness	0.055	0.18	0.28	0.16	0.057	0.072	0.036	0.019	-0.093	-0.056	1

Fig. 3. Pearson Correlation Between LIWC and OCEAN Values

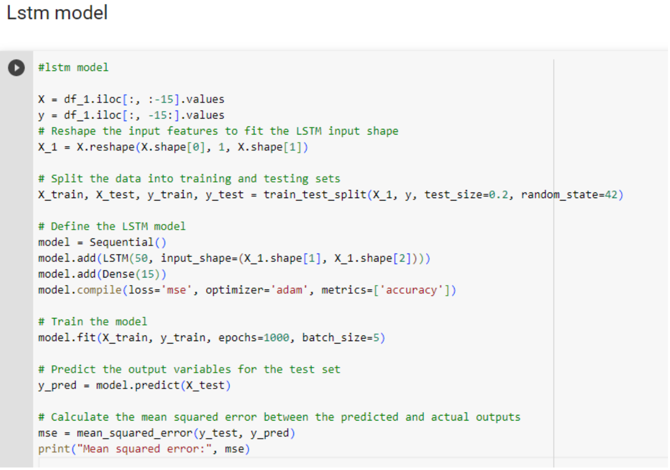
For linear regression, feature selection using Pearson correlation allows us to identify the features that have the strongest linear relationship with the target variables. By selecting these highly correlated features, we can improve the performance of the linear regression model by including the most influential predictors.

Additionally, Pearson correlation can also be useful for feature selection in k-means clustering. It helps identify features that have a significant influence on the clustering results by measuring their correlation with the cluster assignments. By selecting features with high correlation values, we can focus on the variables that are most relevant for clustering the data effectively.

In summary, Pearson correlation provides valuable insights for feature selection in both linear regression and k-means clustering, helping us identify the most influential features related to the Big Five personality traits and Schwartz values.

C. Models Used & Architecture

1) LSTM (Long Short-Term Memory): The LSTM (Long Short-Term Memory) model is a type of recurrent neural network (RNN) that is particularly effective for sequence-based data analysis, including regression problems. It can capture and learn complex temporal dependencies in the input data, making it suitable for tasks such as time series forecasting or regression tasks where the input features have sequential or temporal patterns. For a regression problem, the goal is to predict a continuous target variable. In this case, the LSTM model can be used to learn the relationship between the input features and the target variable and make predictions.

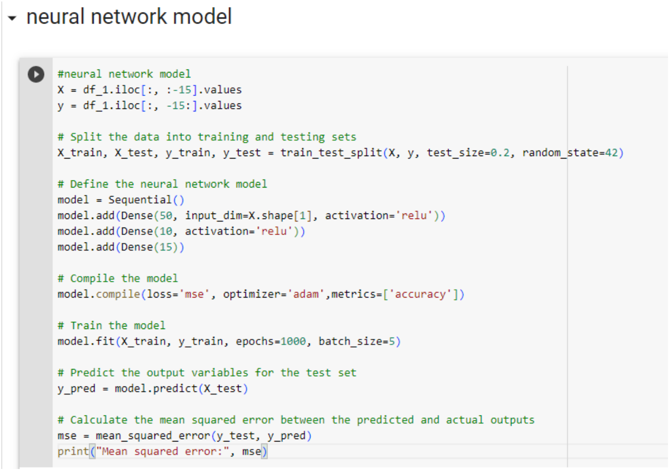


Lstm model

```
#lstm model

X = df_1.iloc[:, :-15].values
y = df_1.iloc[:, -15:].values
# Reshape the input features to fit the LSTM input shape
X_1 = X.reshape(X.shape[0], 1, X.shape[1])

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X_1, y, test_size=0.2, random_state=42)

# Define the LSTM model
model = Sequential()
model.add(LSTM(50, input_shape=(X_1.shape[1], X_1.shape[2])))
model.add(Dense(15))
model.compile(loss='mse', optimizer='adam', metrics=['accuracy'])

# Train the model
model.fit(X_train, y_train, epochs=1000, batch_size=5)

# Predict the output variables for the test set
y_pred = model.predict(X_test)

# Calculate the mean squared error between the predicted and actual outputs
mse = mean_squared_error(y_test, y_pred)
print("Mean squared error:", mse)
```

Fig. 4. LSTM code Snippet

2) Neural network model: The neural network model for regression is a flexible and powerful approach for solving regression problems. It can handle multiple output variables, making it suitable for cases where you need to predict multiple continuous values simultaneously.



neural network model

```
#neural network model
X = df_1.iloc[:, :-15].values
y = df_1.iloc[:, -15:].values

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Define the neural network model
model = Sequential()
model.add(Dense(50, input_dim=X.shape[1], activation='relu'))
model.add(Dense(10, activation='relu'))
model.add(Dense(15))

# Compile the model
model.compile(loss='mse', optimizer='adam',metrics=['accuracy'])

# Train the model
model.fit(X_train, y_train, epochs=1000, batch_size=5)

# Predict the output variables for the test set
y_pred = model.predict(X_test)

# Calculate the mean squared error between the predicted and actual outputs
mse = mean_squared_error(y_test, y_pred)
print("Mean squared error:", mse)
```

Fig. 5. Neural network model code Snippet

D. Model Evaluation

MSE stands for Mean Squared Error, which is a commonly used metric to evaluate the performance of regression models. It measures the average squared difference between the predicted values and the actual values in a regression problem. MSE is a non-negative value, with a lower MSE indicating a better fit of the regression model to the data. A MSE of 0 would indicate a perfect prediction where the predicted values exactly match the actual values. In the case of Mean Squared Error (MSE), a lower value is better. MSE measures the average squared difference between the predicted and actual values in a regression problem. By squaring the differences, MSE penalizes larger errors more heavily than smaller errors. Therefore, a lower MSE indicates that the predicted values are closer to the actual values on average.

When comparing different regression models or evaluating the performance of a single model, a lower MSE suggests better accuracy and a better fit to the data. It indicates that the model has lower overall prediction errors and is better at capturing the underlying patterns and relationships in the data.

LSTM
Mean squared error: 0.04378527921900507

Trait	Actual Output	Predicted Output
Benevolence	0.9	0.8485898
Universalism	0.8	0.90760005
Self-direction	0.9	0.9471413
Stimulation	1.0	0.72462326
Hedonism	1.0	0.5741969
Achievement	0.9	0.64711475
Power	0.6	0.74457395
Security	0.8	0.89032745
Conformity	0.7	0.69685876
Tradition	0.9	0.91032803
Openness	0.575	0.7606051
Conscientiousness	0.5	0.5292755
Extraversion	0.594	0.53294194
Agreeableness	0.75	0.5842514
Neuroticism	0.531	0.52611405

Mean squared error: 0.03300769961871702

Trait	Actual Output	Predicted Output
Benevolence	0.9	0.8313
Universalism	0.8	0.72797
Self-direction	0.9	0.816
Stimulation	1.0	0.719
Hedonism	1.0	0.592
Achievement	0.9	0.707
Power	0.6	0.717
Security	0.8	0.871
Conformity	0.7	0.692
Tradition	0.9	0.829
Openness	0.575	0.63459046
Conscientiousness	0.5	0.51003
Extraversion	0.594	0.53275
Agreeableness	0.75	0.55075
Neuroticism	0.531	0.49582

Neural network model
Mean squared error: 0.037709386297901895

Trait	Actual Output	Predicted Output
Benevolence	0.9	0.8266574
Universalism	0.8	0.705155
Self-direction	0.9	0.9560528
Stimulation	1.0	0.6620714
Hedonism	1.0	0.84992284
Achievement	0.9	0.77654254
Power	0.6	0.8136946
Security	0.8	0.89463913
Conformity	0.7	0.65855753
Tradition	0.9	0.8500715
Openness	0.575	0.7100212
Conscientiousness	0.5	0.4866228
Extraversion	0.594	0.49150893
Agreeableness	0.75	0.5989796
Neuroticism	0.531	0.54879653

For pearson correlation based feature selection

LSTM
Mean squared error: 0.026545864491735936

Trait	Actual Output	Predicted Output
Benevolence	0.9	0.8559785
Universalism	0.8	0.8012208
Self-direction	0.9	0.8031674
Stimulation	1.0	0.76268375
Hedonism	1.0	0.7316604
Achievement	0.9	0.6909019
Power	0.6	0.5399648
Security	0.8	0.802771
Conformity	0.7	0.73848414
Tradition	0.9	0.81720215
Openness	0.575	0.6215603
Conscientiousness	0.5	0.59876645
Extraversion	0.594	0.4979372
Agreeableness	0.75	0.6410849
Neuroticism	0.531	0.46835318

Trait	Actual Output	Predicted Output
Benevolence	0.9	0.8801
Universalism	0.8	0.72992503
Self-direction	0.9	0.769
Stimulation	1.0	0.587
Hedonism	1.0	0.701
Achievement	0.9	0.747
Power	0.6	0.703
Security	0.8	0.83518
Conformity	0.7	0.71532877
Tradition	0.9	0.815
Openness	0.575	0.59793092
Conscientiousness	0.5	0.58219
Extraversion	0.594	0.5023274
Agreeableness	0.75	0.63729242
Neuroticism	0.531	0.44621993

Neural Network Model:
Mean squared error: 0.02624698829769021

Trait	Actual Output	Predicted Output
Benevolence	0.9	0.862
Universalism	0.8	0.8
Self-direction	0.9	0.75
Stimulation	1.0	0.798
Hedonism	1.0	0.854
Achievement	0.9	0.902
Power	0.6	0.78
Security	0.8	1.025
Conformity	0.7	0.817
Tradition	0.9	0.993
Openness	0.57	0.635
Conscientiousness	0.5	0.579
Extraversion	0.59	0.45
Agreeableness	0.75	0.657
Neuroticism	0.53	0.557

IV. Conclusion

In this study, applying a dimensionality reduction technique to the LIWC dataset resulted in improved clustering outcomes. The reduced-dimensional representation allowed for better separation and grouping of data points based on their inherent patterns and similarities. This indicates that the underlying structure of the LIWC dataset was effectively captured and utilized for clustering purposes, leading to enhanced clustering results.