Personality analysis using NLP

1st Mahmuda Junainah
Dept. of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
mahmuda.Junainah@g.bracu.ac.bd

2nd Aurchi Roy
Dept. of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
aurchi.roy@g.bracu.ac.bd

3rd Md. Ehtesham-ur-rahman aurid
Dept. of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
ehtesham.ur.rahman@g.bracu.ac.bd

4th Readhwana Reaz Adrin
Dept. of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
readhwana.reaz.adrin@g.bracu.ac.bd

5th Rifah Tasnia
Dept. of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
rifah.tasnia@g.bracu.ac.bd

6th Annajiat Alim Rasel
Dept. of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
annajiat@gmail.com

7th Md Humaion Kabir Mehedi
Dept. of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
humaion.kabir.mehedi@g.bracu.ac.bd

8th Abid Hossain
Dept. of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
abid.hossain@g.bracu.ac.bd

***Abstract*—The combination of natural language processing (NLP) and machine learning has affected many aspects of our lives in our ever-changing digital landscape. NLP often subtly affects our relationships in everything from voice commands to complex online interactions. Our study investigates personality prediction based on textual analysis of interest in this context. Our study uses advanced datasets and methods, such as polynomial naive Bayes, recursive neural networks, convolutional neural networks, random forests, support vector machines (SVM) Our modeling process starts with painstaking data preprocessing , where transcription reduces noise by removing stopwords The validation loss and accuracy plots shown clearly show the evolution of the model and provide information about its learning process. Furthermore, a thorough analysis using the illusion matrix reveals the robustness of model performance. The predictive ability of the models is demonstrated through measures of precision, recall, and F1-score, and reflects the process of extracting personality traits from text. Our research essentially bridges the gap between linguistics and machine learning, encouraging readers to explore the fascinating interplay between language and personality prediction. This discovery provides a way to better understand the patterns in our language as the digital age advances**

***Index Terms*—Natural Language Processing (NLP), Text Analysis, Sentiment Analysis, Personality Prediction, Machine Learning Models, Recurrent Neural Networks (RNN), Convolutional Neural Network (CNN), Supervised Machine Learning, Naive Bayes, Random Forest, Support Vector Machine (SVM), Data Preprocessing, Model Training**

I. Introduction

Nowadays, natural language processing has become an effective tool not just in theoretical science but also in practical life. From voice typing to text auto correction, along with google translator, everywhere we are using nlp without even realizing. Ai chatbots can not function or communicate with the user without pre-programmed nlp. Moreover, as mankind is moving fast towards globalization and digitalization, we have to use smart devices a lot. For example, in the past , the works which were done offline with pen and paper are now being done by computer with texts. Moreover, almost every adult is now somehow connected to any kind of social media. Social media is not just being used for personal reasons. It is also being used for professional reasons such as marketing, building connections, virtual meetings etc. Therefore,as a human nowadays spends a noticeable amount of time texting, it is a possibility to predict a person's sentiment or personality by analyzing texts, using nlp and machine learning models. In order to make our research a success, we used a large dataset from kegel, and various machine learning models were used in order to increase accuracy. Lastly, we have a vision to extend our research and add more versatile features to our work.

II. Detailed Literature Review

A. Behavioral Mapping, Using NLP to Predict Individual Behavior: Focusing on Towards/Away Behavior

For this paper, the authors collected the data using questionnaires and data provided by the experts. Later on, they applied five different methods. These are IBM emotion analysis, ACS sentiment analysis, vector-based method, logistic regression, and lastly bag-of-words method. Then five models were used during this analysis which are logistic regression (LR), Random Forest Classifier (RFC), Multinomial Naive Bayes (MNB), Complement Naive Bayes (CNB), and Gaussian Naive Bayes (GNB). Among all of them, IBM emotional analysis predicted the unseen data better than other methods although it has the lowest AUC score and accuracy. The average sentiment score for toward data is 0.78 and for away, it is -0.36. However, the study is limited due to its lack of data and imbalanced data, as well as time-related constraints that led to incomplete labeling of words in the datasets.

B. Machine Learning Based Personality Classification for Car Pooling Application

In this paper, the authors used machine learning models (ML) to identify customer personality types from tweets and group them with similar personalities for carpooling. The objective of this study was to group individuals with similar personalities in the same vehicle, bridge the gap between people who are uncomfortable in traveling alongside those with different personalities, and make carpooling easier. In this paper, the authors used the Myers-Briggs Personality Type dataset from Kaggle and preprocessed the dataset using different NLP techniques before feeding the dataset data into the machine learning models. The preprocessing techniques used by the authors include removing unnecessary contents such as removing URL, '', '@', removing all non-characters, etc, encoding each personality category using Sklearn, LabelEncoder and then applying the different NLP techniques such as tokenization, stopping words removal, stemming and lemmatization. After the data was preprocessed, the authors passed the preprocessed data into the ML models to predict the personality of the individuals. The ML algorithms that were used in this paper are XG Boost, Decision Tree Classifier, Support Vector Machine, and Stochastic Gradient. The output of the machine learning algorithms was passed into a matching algorithm and vectorized using a count vectorizer, then the vectorized text was again passed into the pre-trained machine learning models to predict the personality type of the individuals. Finally the output was stored in a database. In this paper. The XG Boost ML model outperformed the other ML models by achieving the highest accuracy of 68%. However, the limitation of this study is that the proposed method may not be optimal for small customer numbers with diverse personality types.

C. Analysis of news sentiments using natural language processing and deep learning

In this paper (Analysis of news sentiments using natural language processing and deep learning) author focuses on analyzing text using nlp and predicting public sentiments regarding news, especially financial news related to the stock market . This helps businessmen a lot as market sentiments influence information flow and trading, thus trading firms hope to profit based on forecasts of price trends influenced by sentiments in financial news (Ruiz-Martínez et al. 2012). The author used multiple datasets, especially from the DJIA Database, which is from Kaggle and contains 25 daily articles with financial news from 2008 to 2016 that the author pulled from the most popular articles on Reddit WorldNews . After pre-processing the datasets with nlp , the author used a deep learning function that imitates the mechanisms of the human brain for finding patterns. The function is :- $(y \log(p) + (1 - y) \log(1 - p))$. Lastly, it is one of very few papers closely related to our topic and helped us in improving our work.

D. A Framework for Automated Personality Prediction from Social Media Tweets

In this paper, the authors propose a framework that employs social media data to predict an individual's personality using the Big Five personality model. This model, also known as the Five Factor Model, encompasses five broad personality traits: openness, conscientiousness, extraversion, agreeableness, and neuroticism. The authors utilize Natural Language Processing (NLP) and the Natural Language Toolkit (NLTK) along with Long Short-Term Memory (LSTM) neural networks to analyze the sentiment present in social media comments and predict the Big Five personality traits. The study finds that personality traits can be accurately predicted from users' tweets, achieving an accuracy of 92%. The proposed methodology consists of three main steps: data collection, data pre-processing, and model building. The authors utilize a dataset of approximately 14,183 tweets labeled with 11 parameters related to emotions, such as anger, joy, fear, and trust. Unfortunately, the paper does not provide explicit information about the source or origin of the dataset. NLP and NLTK are employed for pre-processing the dataset, followed by training the LSTM model to predict the Big Five personality traits.The authors highlight that the relationship between text content and personality traits, as well as the inclusion of contextual information from images, can contribute to accurate personality prediction. The study shows that certain emotions and sentiments are closely associated with specific personality traits. The LSTM model's accuracy surpasses that of other techniques, indicating its capability to accurately predict user personality traits. By utilizing NLP, sentiment analysis, and LSTM neural networks, the authors achieve a high accuracy rate in predicting Big Five personality traits.The authors suggest that this approach could be extended to incorporate more complex interactions and contextual factors for even more accurate predictions.

E. Framework for Implementation of Personality Inventory Model on Natural Language Processing with Personality Traits Analysis

The main focus of the paper is on the importance of textual prediction of personality, highlighting different methods of accurate prediction Once the language used in interviews is established in response to a deeper understanding of the main emphasis, this research goes to the investigation of different patterns and processes that predict personality traits. The article also discusses the dynamic development of personality prediction and highlights its many applications in psychology and computer science. One of the most important studies in this area uses the Big Five framework to build a personality prediction engine using knowledge derived from Facebook interaction data. The study uses techniques such as word processing and N-grams to efficiently examine predictive accuracy across multiple data sets. The study considers the potential implications of these revelations, also moving into areas such as the criminal justice system, and concludes with recommendations for improving current predictive tools The current state of personality prediction based on online data is also discussed in this literature review, emphasizing the need for creative regression algorithms and rich data sets to increase the accuracy of predictions What notably, the study identifies a wide range of potential applications of the criminal internal justice system, from specific integration to computerization The entire article deftly navigates analyzes, emphasizing the importance of predicting identity from data, the process of development, and several promising practical approaches for this young topic

III. Methodology

A. Collected data

For this paper, we have selected the "MBTI Personality Types 500 Dataset." This specific dataset has been collected from www.kaggle.com. The MBTI Personality Types 500 Dataset contains a wealth of insights into the human psyche, centered around the Myers-Briggs Type Indicator (MBTI). This dataset shows how people see the world and make decisions based on their personality types. It consists of around 106,000 preprocessed records, each consisting of posts along with the personality type of the author. These posts have been standardized to be of equal size, containing 500 words per sample. To ensure data quality and uniformity, the following preprocessing steps were applied: (1) Removal of punctuations, stopwords, and URLs, (2) Lemmatization of words and (3) Reconstruction of samples to ensure consistency in size (500 words). The personality types in the dataset encompass a diverse array of 16 unique values, representing the various cognitive functions and preferences individuals possess. The dataset also presents a rich diversity of perspectives, with the following personality type distribution: (1) INTP: 24% (2) INTJ: 21% (3) Other (combining various personality types): 55%. In order to create the dataset , Dylan Storey and Mitchell Jolly collected about 1.7 million posts from Reddit and around 9,000 posts from a forum called PersonalityCafe. This dataset also has a license of ”CC0: Public Domain”. It is basically in csv format named as MBTI 500.csv and has a total of 106,067 records with posts and personality types. It is about 346.05 MB in size. Moreover, the MBTI Personality Types 500 Dataset offers a comprehensive collection of preprocessed posts and personality types, providing researchers with a valuable resource to explore the interplay between cognitive functions, psychological preferences, and the way individuals express themselves online.

B. Proposed methodology

We have preprocessed our dataset by removing stopwords and unnecessary characters such as URL, newline characters and non-alphabetic characters from the dataset. The dataset was randomly shuffled and divided into training and testing set in the ratio of 90% and 10% respectively. Tokenization and label encoding were then applied to the training and testing set. Tokenization involved breaking the paragraphs into smaller units and label encoding involved converting categorical data into numerical data. These modifications allowed the data to be compatible with machine learning models which required numerical input.

C. Proposed Model Architecture

In this paper, we will use different machine-learning techniques to predict personality traits from our dataset. The machine learning techniques that we will use are recurrent neural networks (RNN) or convolutional neural network model (CNN) and some supervised machine learning algorithms such as multinomial Naive Bayes classifier, random forest, and support vector machine (SVM). In our proposed methodology we will start by utilizing the preprocessed dataset as an input for our machine learning models. This preprocessed dataset has undergone steps such as the removal of unnecessary characters and stopwords. We will then train this model using the training set and iterate the data for 10 epochs to improve the performance of the model. Once the training phase is complete, we will save the model and assess the performance of the model through training and validation loss and accuracy graphs. This visualization helps us to understand how well the model is learning during training and whether it is susceptible to overfitting or underfitting. Finally, we will generate a confusion matrix and evaluation metrics. This evaluation process will enable us to analyze and compare the accuracies of different models in predicting personality traits based on the given dataset.

References

[1] S. Lagree and K. W. Bowyer, ”Predicting ethnicity and gender from iris texture,” 2011 IEEE International Conference on Technologies for Homeland Security (HST), Waltham, MA, USA, 2011, pp. 440-445, doi: 10.1109/THS.2011.6107909.

[2] Vicari, M., Gaspari, M. ”Analysis of news sentiments using natural language processing and deep learning.” AI Soc 36, 931–937 (2021). https://doi.org/10.1007/s00146-020-01111-x

[3] P. William, Y. N, V. M. Tidake, S. Sumit Gondkar, C. R and K. Vengatesan, ”Framework for Implementation of Personality Inventory Model on Natural Language Processing with Personality Traits Analysis,” 2023 International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT), Bengaluru, India, 2023, pp. 625-628, doi: 10.1109/IDCIoT56793.2023.10053501.

[4] T. S. Kanchana and B. S. E. Zoraida, ”A Framework for Automated Personality Prediction from Social Media Tweets,” 2022 IEEE World Conference on Applied Intelligence and Computing (AIC), Sonbhadra, India, 2022, pp. 698-701, doi: 10.1109/AIC55036.2022.9848840.