Table of Contents

ABSTRACT

The Russo-Ukrainian conflict has been a highly contentious and protracted geopolitical issue that has garnered significant attention on various social media platforms, particularly Twitter. Online discussion on Twitter has been the main platform for the protracted and polarizing geopolitical conflict between Russia and Ukraine. Because it is a platform that generates a lot of user-generated information, it allows us to investigate the prospect of using tweets related to the dispute as a method to predict people's personality traits using Twitter posts related to the Russo- Ukrainian conflict in this study.

This study uses machine learning algorithms and natural language processing (NLP) techniques to extract and analyze textual data from tweets about the Russia-Ukraine conflict. The main goal of the study is to predict the personalities of Twitter users based on the information they posted during the conflict.

We obtained the dataset from the internet via Kaggle, then pre-processed the data and applied linguistic characteristics on the data using these characteristics and features, we predicted the Big Five scores and personality traits. The accuracy of these personalities is achieved by the machine learning and deep learning algorithms such as Support vector machine (SVM), MLP (Multilayer Perceptron), and RCNN (Region based convolutional neural network).

1.INTRODUTION

Along with the rise of social media, recent years have seen an exponential growth in information, particularly in the form of textual data types. As of January 2023, there were 4.8 billion active social media users globally. People commonly use social media to express opinions on issues related to politics, psychology, money, relationships with others, and the environment. They also use it to discuss their daily lives and the welfare of their families. Sometimes people will use these expressions to describe the behavior and characteristics of others. In fact, prior study demonstrates a strong association between social media user personalities and their online behavior.

The study of personality using tweets from Twitter has gained popularity among NLP researchers in recent years. Researchers now have access to a plethora of data on people's online behavior thanks to the growing popularity of social media platforms, which may be used to spot trends and patterns in personality traits. Studies reveal that social media profiles can properly predict extraversion and neuroticism among other personality traits. The personality can be described as a style of influence or utilized to distinguish individual persons. Marketers that wish to target clientele groups or psychologists who want to comprehend how personality traits are portrayed online may find this information beneficial. But it is important to remember that social media profiles only provide a brief glimpse into someone's personality and should not be relied upon as the only source of information.

Numerous psychological analyses were undertaken in try to predict personality. The language people use in tweets can be used to infer a person's level of extraversion, openness, agreeableness, conscientiousness, and emotional stability. Based on the findings, numerous academics created a variety of models that enable identifying the traits that determine an individual's personality. By these models, connections between personality and psychological diseases, job satisfaction and performance, and even interpersonal interactions can be found. Because they contain a wealth of information and have millions of users, social networks are a fantastic place to start when researching the personalities of groups.

In the study under consideration, a personality test is employed to forecast the temperaments of Twitter users. Several publications have made use of neural networks and the Twitter user personality test. The current personality prediction system used open vocabulary feature extraction with a deep learning and machine learning method to improve classification accuracy.

2.BACKGROUND AND RELATED WORK

2.1 The Big Five Model

Using information from Twitter, one may predict personalities for a while. The Big five personality model is used to map a variety of research job status data and personality traits. The openness, conscientiousness, extraversion, agreeableness, and neuroticism elements of the Big Five model of personality include openness, conscientiousness, extraversion, agreeableness, and neuroticism.

- **Openness:** A person's level of receptivity to new ideas and experiences is referred to as their level of experiential openness. High openness individuals are inquisitive, creative, and appreciative of diversity.
- **Conscientiousness:** Conscientiousness is a measure of how well-organized, dependable, and diligent a person is. High conscientiousness individuals are frequently adept at goal-setting and carrying them out.
- **Extraversion:** The degree to which a person is outgoing and enjoys interacting with others is referred to as extraversion. Extraverted folks often get energized by social settings and like being around other people.
- **Agreeableness:** How cooperative, helpful, and trustworthy people are depends on how agreeable they are. Agreeableness is the degree to which people are prone to positive emotions, such as happiness and agreeableness is the extent to which people are prone to negative emotions, such as anxiety, wrath, and sadness.
- **Neuroticism:** High neurotic individuals are frequently more sensitive to stress and more likely to feel negative emotions.

It has been established that the Big Five model is a genuine and trustworthy personality assessment tool. It has been applied in a wide range of research projects, including ones looking at relationships, mental health, and work performance. To better understand ourselves and others, the Big Five paradigm can be useful. The Big Five model has been shown to be a reliable and valid measure of personality. It has been used in a wide variety of research studies, including studies of job performance, mental health, and relationships. The Big Five model can be a helpful tool for understanding ourselves and others better.

2.2 Applications of the Big Five

Numerous human behaviors and decisions have been found to be associated with the Big Five personality traits. Preferences like music, politics, and pet ownership all relate to personality. Big Five personality traits have been proven to reliably predict a consumer's choice for national brands or independent brands in the context of marketing and advertising. It has been demonstrated that personality traits have an impact on managerial behavior, team performance, unproductive behaviors, entrepreneurial status, and job performance and satisfaction in the workplace.

It has been demonstrated that personality and interface preference are related in the field of human-computer interaction. Users preferred a graphical user interface that matched their personality type and were more inclined to purchase a book that included reviews written from that personality type's point of view.

According to these studies, it is possible to use the Big Five personality traits to better understand and forecast people's behavior in a range of situations. This knowledge can be applied to better team management, staff recruitment, and product and service design.

2.3 Importance of Personality prediction in social media

Due to its potential impact on numerous applications and areas, personality prediction in social media has attracted a lot of attention. For personalized user experiences, targeted advertising, content recommendation, and social media analytics, it is essential to understand a person's personality traits since they can offer important insights into their behavior, preferences, and habits. Platforms can give specialized content and recommendations that match user's tastes and features by successfully forecasting personality traits from social media data. Improved user experiences and higher platform usage result from this personalized strategy, which also increases user engagement and happiness.

Additionally, the ability to anticipate personality can help with targeted advertising campaigns that appeal to user's personality types. Personality-specific advertisements have a higher chance of grabbing people's attention, evoking favorable reactions, and achieving intended results. This increases user pleasure by lowering obtrusive and irrelevant advertising, which not only benefits advertisers by maximizing their marketing efforts. Social media analytics and research are also made easier by personality prediction in social media. Researchers can find patterns, trends, and correlations between different social phenomena and personality qualities by analyzing large-scale social media data with personality insights. These discoveries advance our knowledge of social interactions, psychological processes, and human behavior, which makes it possible for social scientists to create more precise models and hypotheses.

The Ukrainian-Russian conflict may have an impact on people's personalities, as seen in their tweets, but few research have particularly examined this possibility. It might be feasible to find patterns that reveal not just what people say about geopolitical events but also who they are likely to be based on those comments by examining tweet content pertaining to the dispute among the users.

This kind of analysis could ultimately help us better understand not only predictors but also mental health issues like anxiety and depression that could potentially affect many people during periods like these where tensions are high around international conflicts. This would support future interventions that are more effective.

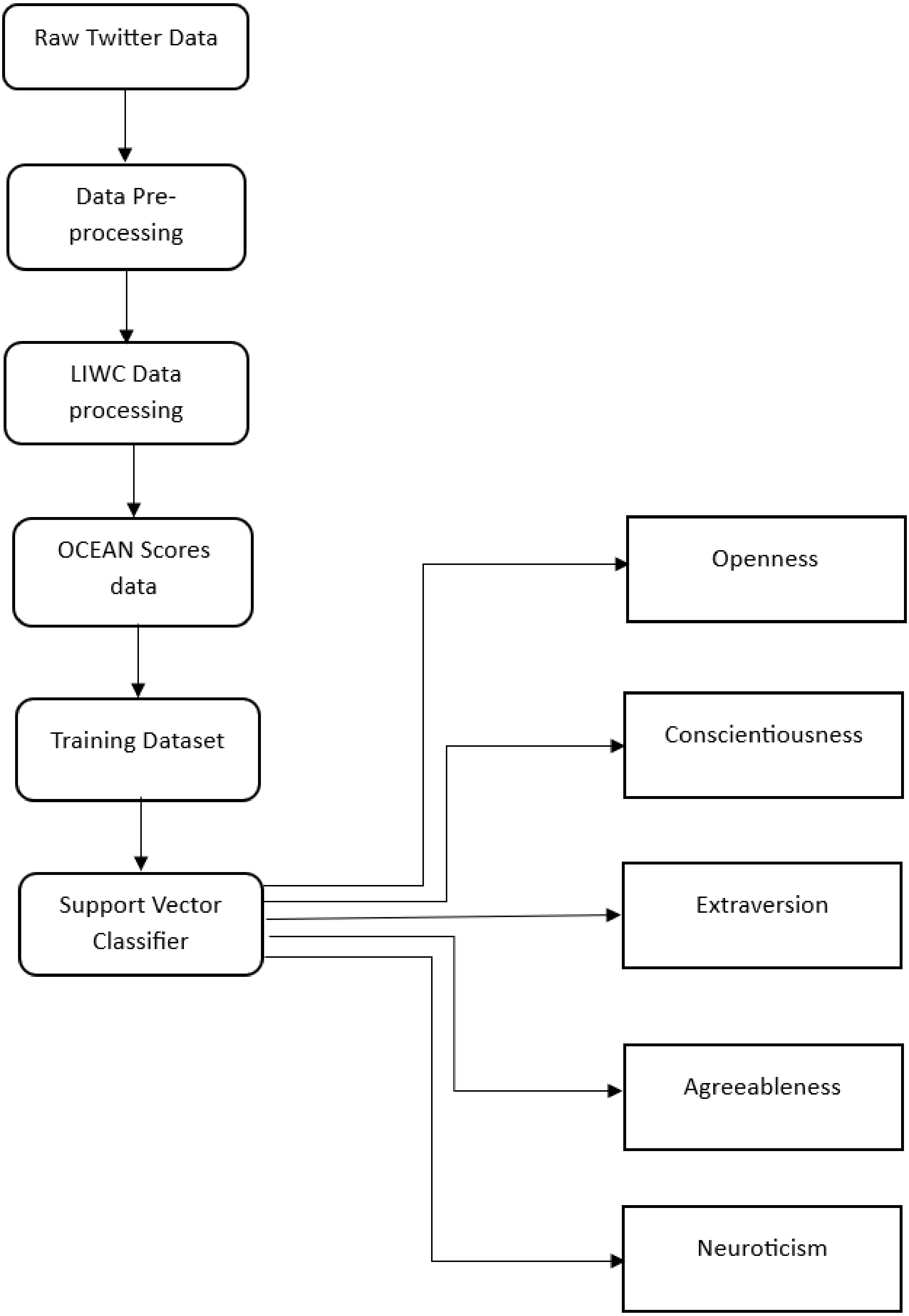
3.Methodology

The gathering of data is the first step in the research cycle, followed by analysis of that data and use of pre-processing techniques to produce textual data. From the textual information of the tweets, the OCEAN Scores can be found. Some machine learning and deep learning algorithms employ the OCEAN Scores to forecast the user's personality, and they also evaluate the accuracy and metrics of the results. Flow Chart

3.1 Data Collection

The dataset used in this study has 47,994 rows and 29 columns of data that indicate different attributes related to each tweet. The first column, “Unnamed: 0”, seems to be an unnamed column that could be the index or a special identification for each row. The “username” column carries the Twitter account's username, while the “userid” column contains the user ID of the Twitter account connected to the tweet. The “acctdesc” column contains the Twitter account's bio or account description, while the 'location' column shows the address listed in the account profile.

Other columns include “following,” which shows the number of accounts the user is following, and “followers,” which shows the total number of followers the person has. The total number of tweets produced by the user are listed in the column labelled “totaltweets,” and the timestamp at which the user's account was formed is listed in the column labelled “usercreatedts.” The “tweeted” column displays the specific ID for the tweet, and the “tweetcreatedts” column includes the timestamp for its creation.


Raw Twitter Data
Data Pre-processing
LIWC Data processing
OCEAN Scores data
Training Dataset
Support Vector Classifier
Openness
Conscientiousness
Extraversion
Agreeableness
Neuroticism

The number of times the tweet has been retweeted is indicated by other columns like "retweetcount," while the "text" column carries the tweet's text. The "language" column denotes the language of the tweet, and the "hashtags" column specifies any hashtags that were used in it. The "coordinates" column lists the tweet's associated geographic coordinates.

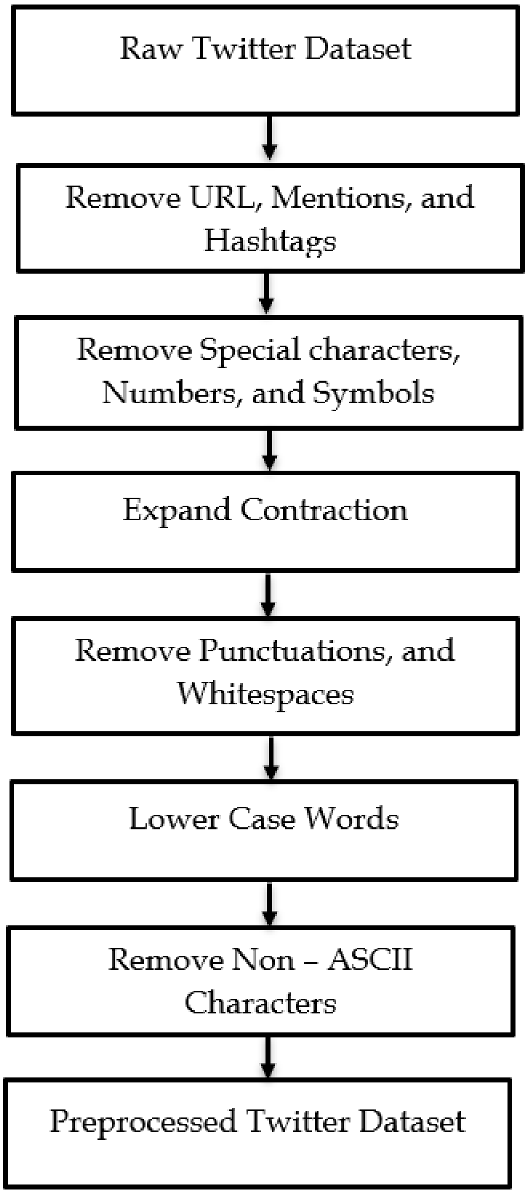
Additional columns include "favorite_count" (the number of times the tweet has been favorited), "is_retweet" (which indicates whether the tweet is a retweet or an original tweet), "original_tweet_id" (the ID of the original tweet if the current tweet is a retweet), and related columns for the original tweet's user ID and username. The dataset also includes columns to manage replies, quoted tweets, and the IDs, user IDs, and usernames that go along with them.

The timestamp when the tweet was extracted or collected is displayed in the "extractedts" column. We focus on relevant elements that contribute to the prediction and categorization of personality traits on social media platforms when doing the necessary analytical and modelling activities for this study, which involves selectively using attributes from this dataset.

3.2 Data Pre-Processing

The pre-processing stage involves performing different data cleansing and transformation operations on the dataset's chosen columns. Three distinct columns in this study project are picked for more examination. "serial_no", the first column, is used as a serial number or unique identifier for each entry in the collection. This column assists in indexing and organizing the data for simpler tracking and reference. "Username," the second column, displays the username connected to each tweet. The username reveals the identity of the tweet's author and can be used to spot trends or personality traits associated with users. The exact text of each tweet is found in the third column, which is labelled "tweet". The focus of the analysis is on this column because the textual data from the tweets is essential for gaining insights and conducting additional research. Flow Chart

Typically, the pre-process is applied to the dataset. The use of URLs, symbols, mentions, hashtags, special characters, numerals, punctuation, excess whitespace, and emoticons in social media status will be deleted from all previously gathered data. the use of must've to become must have been an example of expanding a contraction in a sentence. The next step is to lowercase each sentence to make it more normal. Additionally, to avoid ambiguity, any stop words and clitics will be eliminated. The re and string libraries, which offer several linguistic functions to


Raw Twitter Dataset
Remove URL, Mentions, and Hashtags
Remove Special characters, Numbers, and Symbols
Expand Contraction
Remove Punctuations, and Whitespaces
Lower Case Words
Remove Non – ASCII Characters
Preprocessed Twitter Dataset

help with cleaning up social media status data, are used to pre-process the data as part of this procedure.

To translate from several languages to solely English language, there will be an additional step during Twitter data pre-processing. In this study, "Lang detect" library is used to carry out this operation. With the use of this method, we can produce a consistent and homogenous dataset for the analyses that follow, enabling us to find patterns, trends, and insights that are unique to the English language environment. We can effectively explore the connection between linguistic patterns and personality traits by limiting the dataset to English tweets. This allows us to gain insightful knowledge about how people express themselves and display distinctive traits within the English-speaking social media landscape.

3.2.1 Investigation of the LIWC Tool

The text analysis programme LIWC (Linguistic Inquiry and Word Count) was created by James W. Pennebaker and his associates. The linguistic and psychological characteristics of texts are intended to be analysed and quantified. By using a predetermined dictionary to categorise words into various linguistic and psychological categories, LIWC enables researchers to look at word usage patterns in texts and make assumptions about the psychological and social traits of the author.

Here are some key aspects and uses of LIWC in research work:

1. Linguistic Categories: LIWC offers a wide range of linguistic categories, such as function words (such as pronouns, articles), content words (such as nouns, verbs), social words (such as family, friends), affective words (such as positive emotions, negative emotions), cognitive words (such as insight, certainty), and many others. With the use of these categories, researchers can examine the distribution and usage patterns of terms in each category and capture many facets of language use.

2. Psychological Categories: In addition to linguistic categories, LIWC also has psychological categories that reflect various psychological facets. These categories include feelings (such as joy or sorrow), social dynamics (such as kinship or power), cognitive dynamics (such as causation or insight), and individual concerns (such as one's health or job). Researchers can learn more about the psychological states and contexts by analysing the term's presence and usage in these categories.

3. Text Analysis: LIWC offers several types of text data analysis. It can be used to determine the percentage of words in various categories within a text, gauge the text's emotional tone, look at changes in language use over time, contrast texts from various sources or groups, and pinpoint linguistic markers linked to psychological traits or behaviours.

4. Psychological Insights: The LIWC method has been used in several fields, including psychology, the social sciences, marketing, and communication studies. It has been applied to research language patterns, social dynamics, cognitive processes, emotional expressiveness, and personality traits in a variety of circumstances. Researchers can better comprehend the underlying meaning of text data by using LIWC analysis, which can offer insightful information about the psychological and social elements of text data. Using LIWC, we can extract linguistic features from the Twitter dataset for our studywork, including the distribution of various word categories, emotional tone, cognitive processes, and social ties. Applying LIWC analysis can help comprehend language use's relationships with personality characteristics and other psychologicalfactors in the context of social media data. This analysis strengthens your capacity tolook for relationships between language and personality patterns and associations, which advances the general goals of the study.

3.2.2 Analyzation of the Dataset using LIWC Tool

The LIWC-22 system consists of software and a "dictionary," or a chart that links significant psychological theories and constructs with words, expressions, and other linguistic constructions. Target terms are used to eliminate ambiguity when referring to words found in texts that LIWC22 reads and analyses. The term "dictionary words" will be used to describe words found in the LIWC-22 dictionary file.

To identify the words that are most frequently used in a new data collection, it might occasionally be helpful to look at a word frequency table.

When we input our dataset into the LIWC - 22 software, it pre-processes the data and outputs it in the necessary format.

3.2.3 Obtaining OCEAN Scores using LIWC Tool

Based on prior research that examined the relationship between language use and personality traits, the linguistic variables that were chosen to be employed in computing the OCEAN scores were chosen. These characteristics have been proven to offer insights into people's linguistic patterns and behaviours, which can be a good indicator of their personality characteristics.

Let us look at the traits' chosen qualities in more detail:

- Openness:
  Several characteristics are considered while determining an individual's openness. These include the average number of words per sentence (WPS), which reflects complexity, and the tone of the text, which reveals the emotional quality imparted. BigWords and "cogproc" features track the cognitive processes involved in language use, whereas "BigWords" measures the frequency of employing complicated or unusual terms. Furthermore, features like "insight," "cause," "discrep," "tentat," "certitude," "differ," and "memory" offer guidance on how to express insights, describe causal relationships, talk about contradictions or inconsistencies, express uncertainty or certainty, express differences or opposing viewpoints, and refer to past experiences or personal memories, respectively. The understanding of linguistic complexity, reflection, conceptual thinking, and openness to many points of view are all provided by these linguistic characteristics taken together.

- Conscientiousness:
  It is determined by language characteristics. While "Clout" encapsulates linguistic clout or assertiveness, "Analytic" indicates analytical reasoning and precision. Insights on the overall tone and sentence length are provided by the "Tone" and "WPS" capabilities. Features like "BigWords," "cogproc," "cause," "certitude," "differ," and "work" examine how complex or uncommon words are used, cognitive processes involved in language use, describing causal relationships, expressing certainty or strong beliefs, expressing differences, or opposing viewpoints, and referencing work-related activities, in that order. Precision, aggressiveness, and the language's emphasis on issues of the workplace can all be understood by using these characteristics.

- Extraversion:
  A variety of language factors are taken into consideration while assessing extraversion. The terms "Clout" and "Tone" refer to linguistic impact or assertiveness, respectively, while "Tone" describes the text's overall tone.

  Sentence length and the use of complicated words are captured by the "WPS" and "BigWords" metrics. The "pronoun" feature counts how often pronouns are used, which might be a sign of social orientation or self-reference. Features like "Social," "affiliation," "achieve," "cogproc," "insight," "cause," "discrep," "tentat," "certitude," "differ," and "memory" offer insights into references to social interactions, expressions of affiliation or belonging, discussions of achievements or goals, cognitive processes, expressing insights or understanding, describing causal relationships, describing inconsistencies or contradictions, discussing uncertainty or certainty, and expressing differences. These characteristics can be examined to gain knowledge about a person's assertiveness, engagement, and social orientation.

- Agreeableness:
  Agreeability is measured using a variety of language features. The overall tone of the text is represented by the "Tone" feature. Insights into sentence length and the use of complicated words are provided by "WPS" and "BigWords". The "pronoun" feature counts the number of pronouns used, which can be a sign of self-reference or social orientation. "Social," "affiliation," "cogproc," "insight," "cause," "discrep," "tentat," "certitude," "differ," "memory," "socbehav," "prosocial," "polite," "conflict," "comm," "socrefs," "family," "friend," "female," and "male" features are used to capture references to social interactions, expressions of belonging or affiliation, cognitive causal connections, discussing contradictions or inconsistencies,

expressing scepticism or confidence, expressing differences or opposing viewpoints, referring to earlier instances or personal memories, being polite, engaging in social behaviour and prosocial behaviour, mentioning conflict or disagreement, talking about communication or dialogue,

- Neuroticism:
  By looking at linguistic elements, neuroticism is assessed. While "WPS" and "BigWords" denote sentence length and the use of complicated words, respectively, "Tone" depicts the overall tone of the text. The "pronoun" feature counts how often pronouns are used, which can indicate self-reference or social orientation. Features like "Affect," "tone_pos," "tone_neg," "emotion," "emo_pos," "emo_neg," "emo_anx," "emo_anger," and "emo_sad," which analyse both positive and negative emotions, with an emphasis on anxiety-related emotions, anger-related emotions, and sadness-related emotions, capture emotional content. References to illness or health-related issues (illness), death or mortality (death), and mental health or psychological states (mental) might shed light on a person's neurotic tendencies and wellbeing worries. Understanding these characteristics can assist one comprehend a person's emotional content, adverse effect, and references to mental health.

Using the OCEAN personality trait scores (Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism), some of the machine learning anddeep learning algorithms can be used to predict the personality of every user.

3.2 Model Selection and Evaluation

3.2.1 Personality prediction using OCEAN Scores.

In this level, we need a different dataset with pre-defined scores for the algorithm's training. Additionally, we use our dataset, which consists of OCEAN scores, as testing data, allowing us to forecast each user's personality using machine learning algorithms like Support Vector Classifier (SVC), Random Forest, and Neural Networks.

We employed the Support Vector Classifier (SVC) algorithm to predict the personalities in our research.

Additionally, we experimented with Gaussian Naive Bayes, Random Forest, and Neural Network algorithms. Among these algorithms the Support Vector Classifier (SVC) algorithm is better at predicting personality traits.

3.2.2 Evaluation of personality predictions.

The personality traits obtained by Support Vector Machine algorithm is evaluated with various machine learning and deep learning algorithms. For the evaluation we split our dataset into two datasets training dataset is for training the model and another testing dataset for testing the model. In our work we used the Gaussian Naive Bayes, Random Forest Classifier, Support Vector Classifier (SVC), MLP (Multilayer Perceptron) classifiers, Long Short-Term Memory (LSTM), GRU (Gated Recurrent Unit), and RCNN (Region-based Convolutional Neural Network).

- **Gaussian Naive Bayes**:

Using the Bayes theorem as a foundation, the Gaussian Naive Bayes classifier is a well-known machine learning technique. It improves on the Naive Bayes method by supposing that the data has a Normal or Gaussian distribution. Since it makes mean and standard deviation estimates simple, this assumption makes the implementation easier to complete. The Gaussian Naive Bayes classifier offers several advantageous features when used to personality analysis based on the Big Five model (OCEAN).

Firstly, the algorithm is known for its simplicity and computational efficiency, making it particularly suitable for large-scale text classification tasks involved in personality analysis. It can efficiently process high-dimensional trait spaces commonly used in this domain, enabling faster analysis of large datasets in real-time or near real-time scenarios. Secondly, despite the "naive" assumption of feature independence, the classifier can still yield meaningful results. While feature dependencies may exist, the algorithm can capture important associations between features and class labels. It identifies relevant traits and their contribution to predicting personality traits, even if there are correlations or dependencies between the features.

Moreover, the Gaussian Naive Bayes classifier exhibits robustness when faced with data that deviates from the strict assumptions of a Gaussian distribution. It can handle skewed or noisy data, making it suitable for real-world scenarios where data may not perfectly adhere to theoretical distributions. Despite potential deviations from the assumptions, the classifier provides valuable predictions and insights into individual personality traits. Additionally, the interpretability of Naive Bayes classifiers is noteworthy. They estimate class probabilities and offer insights into the contribution of each trait to the prediction. The simplicity of the algorithm facilitates a clear understanding of how each trait influences the likelihood of a specific personality trait being present. This interpretability is valuable in personality analysis as it aids researchers and practitioners in gaining a deeper understanding of the relationships between OCEAN scores and personality traits.

Output:

```
Naive Bayes Classifier:
Accuracy: 0.770589809940216
Precision: 0.8673348541225273
Recall: 0.770589809940216
F1-score: 0.7969629966805309
```

- **Random Forest Classifier**:

An ensemble learning technique called the random forest classifier uses several decision trees to aggregate predictions. Its name comes from the notion that it is made up of a "forest" of decision trees, each of which is trained using a different random subset of the data and characteristics. The final forecast is then created by averaging the predictions from several trees. The random forest classifier is used in our work will categorize personality characteristics based on OCEAN scores.

Firstly, the random forest classifier is an ensemble method that leverages the collective wisdom of multiple decision trees. By combining predictions from different trees, the model improves its predictive performance and generalization ability. This ensemble approach reduces the risk of overfitting, which is crucial in personality trait classification tasks where the goal is to accurately generalize to unseen data.

Moreover, the random forest classifier provides valuable insights into feature importance. It allows us to assess the relevance and contribution of different OCEAN scores in determining personality traits. By understanding the importance of specific language traits, we can gain a deeper understanding of the underlying relationships between language use and personality traits. This knowledge aids in interpreting the model's predictions and provides valuable insights into the nature of personality analysis. Additionally, the random forest classifier is well-suited for capturing non-linear relationships between OCEAN scores and personality traits. It excels at learning complex patterns in the data, which is particularly important when dealing with intricate relationships between language traits and personalities. By leveraging the collective predictions of multiple trees, the random forest model can effectively capture and represent these non-linear dependencies, enhancing its ability to accurately predict personality traits.

Lastly, the random forest classifier demonstrates robustness against overfitting. Overfitting occurs when a model becomes overly specialized to the training data and performs poorly on unseen data. By aggregating predictions from multiple trees and introducing randomness in the training process, the random forest model mitigates the risk of overfitting. This ensures that the model's predictions are more reliable and generalizable, providing a trustworthy tool for personality analysis based on OCEAN scores.

Output:

```
Random Forest Classifier:
Accuracy: 0.8023556705630409
Precision: 0.7411248009911064
Recall: 0.8023556705630409
F1-score: 0.76735956962452
```

Support Vector Classifier (SVC):

A supervised machine learning method created especially for categorization issues is the SVC algorithm. There are several factors that contribute to its efficacy in our code. The capacity of SVC to capture intricate nonlinear correlations between OCEAN scores and personality factors is a big benefit of using it. Since there may not be a linear link between personality characteristics and qualities, SVC uses kernel methods to translate the input features into higher-dimensional spaces. This modification improves the precision of personality trait categorization by allowing the computer to recognize complex patterns and limits that exist within the data.

Moreover, the SVC algorithm is well-suited for multi-category classification tasks, which is crucial for our project as we are classifying personality traits into multiple categories. SVC effectively handles categorical variables and can classify instances into their respective personality trait categories, including openness, conscientiousness, extraversion, contradiction, and neuroticism. By considering the distinct separation between classes and utilizing support vectors, SVC provides reliable and accurate classification results.

Output:

```
Support Vector Classifier (SVC):
Accuracy: 0.8599982154010886
Precision: 0.7976813919250675
Recall: 0.8599982154010886
F1-score: 0.8245770930580905
```

- MLP (Multilayer Perceptron) classifier:

Our research, which aims to predict personality traits based on the Big Five Models (OCEAN), heavily relies on the MLP (Multilayer Perceptron) classifier, a kind of artificial neural network. MLP classifiers are very useful in personality analysis jobs because to a number of their benefits.

Various variables acquired from text data, such as word frequency, sentence structure, and emotion ratings, can be used to train MLP classifiers in the context of personality research. The MLP classifier uses these characteristics as inputs and learns to categorize people into various personality trait groups using the training examples that are given.

One key advantage of MLP classifiers is their ability to capture nonlinear relationships between input traits and personality traits. Personality traits are influenced by intricate interactions among various factors, and MLP classifiers excel at identifying and leveraging these complex patterns and correlations. This capability allows the MLP classifier to accurately model the complex nature of personality traits and make reliable predictions based on the input text data. MLP classifiers also have the capability of automatically picking out pertinent characteristics from the input data. By doing so, the requirement for labor- and judgment-intensive manual feature development is removed. In the case of large unstructured text datasets, MLP classifiers autonomously discover and utilize essential features, making the personality analysis process more efficient and accurate.

Moreover, MLP classifiers demonstrate good generalization capabilities, enabling them to accurately predict personality traits for new, unseen instances. This is crucial in personality analysis, as the goal is to make accurate predictions for individuals who were not part of the training data. MLP classifiers can learn from the provided training examples and generalize their understanding to make reliable predictions for new individuals.

Output:

```
MLP Classifier:
Accuracy: 0.8810564825555457
Precision: 0.8453993740263046
Recall: 0.8810564825555457
F1-score: 0.8512236079698261
```

- **Long Short-Term Memory (LSTM):**

The utilization of Long Short-Term Memory (LSTM)in our project proves to be highly beneficial for predicting personality traits based on the provided code. LSTM is a type of recurrent neural network (RNN) architecture that excels in processing sequential data, making it well-suited for tasks where the order of input data is crucial, such as natural language processing (NLP) and time series analysis.

The decision to incorporate an LSTM model in our code stems from several key factors. Firstly, LSTM models are specifically designed to handle continuous data that holds significant sequential information. In the context of personality traits, the order in which these traits appear can provide valuable insights for accurately predicting personality types. By leveraging LSTM's ability to process sequential data, the model can effectively capture the dependencies and patterns present in the input traits.

Furthermore, LSTMs address the challenge of capturing long-term dependencies, which is a common limitation in traditional RNNs due to the vanishing gradient problem. In the context of personality prediction, certain traits may have lasting effects on subsequent traits, and it is crucial to capture these long-term dependencies. LSTMs excel in storing information over longer time steps, enabling them to effectively capture and utilize these dependencies for accurate personality predictions. Additionally, the LSTM model is capable of handling variable input lengths, making it suitable for scenarios where the length of input traits may vary from person to person. This adaptability allows the model to accommodate different input lengths and generate predictions based on the available input sequences, enhancing its versatility in analyzing personality traits.

Moreover, the LSTM model employed in the code facilitates multi-class classification, as demonstrated using a SoftMax activation function in the output layer. This enables the model to predict one of several personality types based on the given input characteristics, allowing for comprehensive personality analysis.

Output:

```
LSTM Model:
Loss: 0.2870079278945923
Accuracy: 0.9143316149711609
Validation Loss: 0.28813108801841736
Validation Accuracy: 0.9200499653816223
```

- **GRU (Gated Recurrent Unit):**

To analyze and forecast personality traits based on the input data, the GRU (Gated Recurrent Unit) model, a form of recurrent neural network (RNN) architecture, is used. The GRU model excels at processing sequential data, which makes it especially well-suited for jobs requiring text, voice, and time series data. Using gate mechanisms that only update the hidden state, when necessary, the model is able to capture long-term relationships within the input sequence while addressing the problem of vanishing gradients.

The GRU model is useful in our research for classifying personality characteristics into many categories. To forecast the related personality trait categories, the dataset comprises of attributes linked to various personality qualities. We successfully capture sequential dependencies using the GRU model and manage variable-length input sequences. The model is built using the Kera library and includes thick layers with the proper activation functions. With early stopping added to avoid overfitting, it is trained on a dataset using training data and assessed using validation data.

Overall, the GRU model is an effective method for examining and predicting personality characteristics using the given code. This assignment is a good fit for it because of its capacity to recognize relationships in sequential data and manage variable-length sequences. We can produce precise predictions based on the input data by utilizing the GRU model's capabilities and gaining insightful knowledge about personality features.

Output:

```
GRU Model:
 Loss: 0.2846258878707886
Accuracy: 0.9167410135269165
Validation Loss: 0.2865751385688782
Validation Accuracy: 0.9149638414382935
```

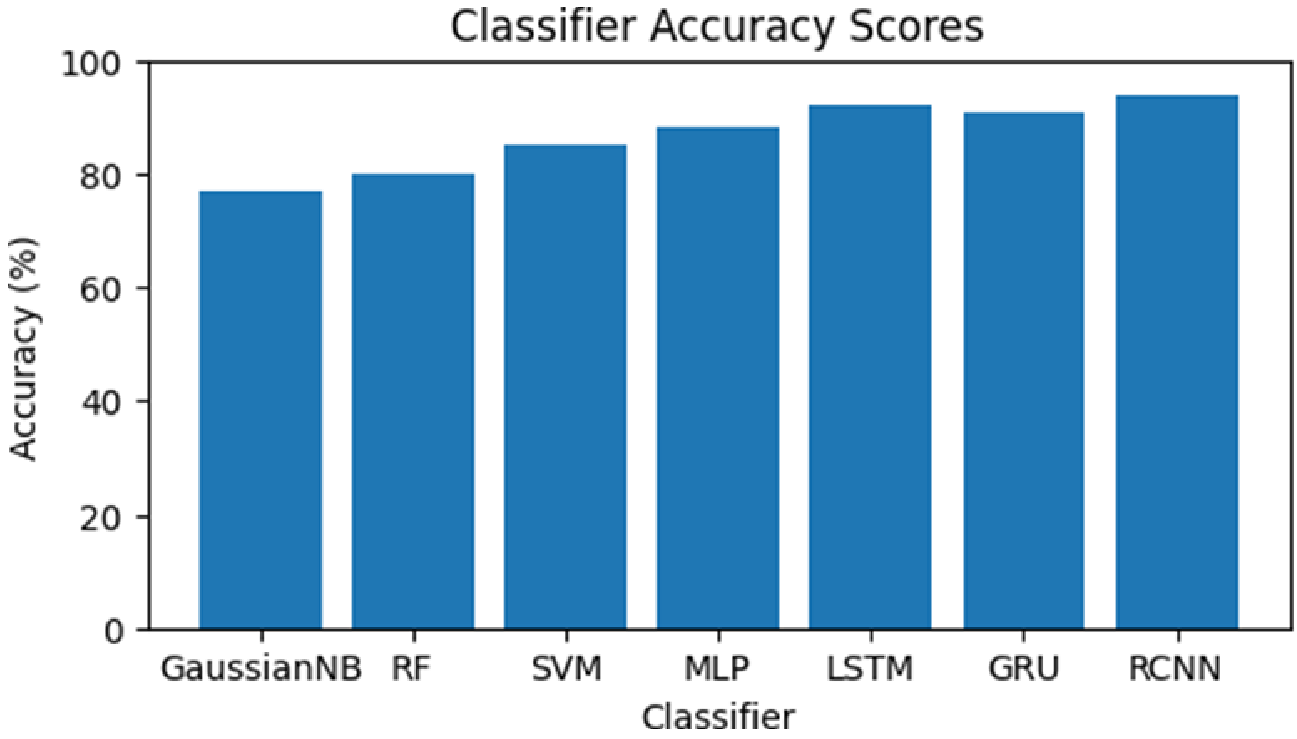
- **RCNN (Region-based Convolutional Neural Network):**

The RCNN (Region-based Convolutional Neural Network) architecture, although primarily designed for image-related tasks, may not directly apply to personality analysis based on OCEAN scores. However, in the field of personality analysis, text data, such as essays and social media posts, is commonly processed. In this context, deep learning models, specifically recurrent neural networks (RNNs), convolutional neural networks (CNNs), and transformer models, play a crucial role in extracting meaningful features from textual data. For the analysis of personality traits based on the OCEAN model, deep learning models excel at capturing linguistic patterns, semantic relationships, and contextual information present in text inputs. By training these models on labeled datasets, they learn to associate specific language patterns and word usage with personality traits, enabling accurate predictions of personality scores for new, unseen text inputs.

Output:

```
RCNN Model
Accuracy: 94.95%
```

- Output of all these models is plotted below.



3.2.3 Evaluation metrics.

- Accuracy:
  When doing classification tasks, accuracy is a standard performance statistic. Out of all the cases in the dataset, the proportion of instances that were properly predicted is what is counted. Overall evaluation of the model's propensity for prediction is given. For precision, use the formula:

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive\ +\ False\ Positive\ +\ True\ Negative\ +\ False\ Negative}$$

3.2.4 Evaluation metrics.

- Accuracy:
  When doing classification tasks, accuracy is a standard performance statistic. Out of all the cases in the dataset, the proportion of instances that were properly predicted is what is counted. Overall evaluation of the model's propensity for prediction is given. For precision, use the formula:

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive\ +\ False\ Positive\ +\ True\ Negative\ +\ False\ Negative}$$

- Precision:
  Precision is the proportion of true positive predictions among all positive predictions generated by the model. It highlights how accurate optimistic forecasts may be. The accuracy is calculated using the formula below:

$$Precision\ = \frac{True\ Positives}{True\ Positives\ +\ False\ Positives}$$

- Recall:
  Recall, sometimes referred to as sensitivity or true positive rate, is a statistic that expresses the proportion of accurate predictions among all real positive cases in the dataset. The emphasis is on compiling all positive instances. Recall is calculated using the formula shown below:

$$Recall = \frac{True\ Positives}{True\ Positives\ +\ False\ Negatives}$$

- F1-score:
  An evaluation of a model's performance is provided by the F1-score, a metric that combines accuracy and recall. It is useful when there is an uneven distribution of students throughout the classes and it is the harmonic mean of recall and accuracy. Following is the formula used to determine the F1-score:

$$F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Conclusion

Finally, the personalities of Twitter users who engaged in the Russian-Ukrainian war were anticipated by our work. The outcomes showed that the RCNN model had the best level of accuracy, coming in at 94%. The Gaussian Naive Bayes algorithm, on the other hand, performed comparatively less accurately, achieving an accuracy of 77%. All results of these models were evaluated with Accuracy, Precision, Recall, and F1-score.

The effect of linguist characteristics derived from the LIWC software showing the efficiency of these qualities was found to be restricted in some situations where the information about the tweet was unclear, even though they offered useful insights into several characteristics. This implies that adding more features or include more contextual data could increase the prediction precision in certain situations.

Future Work

To improve prediction accuracy, sentiment analysis and linguistic characteristics may be included in future personality prediction analysis research. Additionally, a more complex understanding of personality features may result from extending the collection. The most pertinent elements for a precise forecast of personality may be found using feature selection algorithms. Additionally, using these prediction approaches for applications that are specialized to a certain area might produce insightful results. We can progress the study of sentiment analysis-based personality prediction by taking these factors into account.

References

1. Artificial Intelligence Model for the Identification of the Personality of Twitter Users through the Analysis of Their Behavior in the Social Network by William Villegas-Ch. https://www.mdpi.com/2079-9292/11/22/3811
2. Text based personality prediction from multiple social media data sources using pre-trained language model and model averaging by Hans Christian. https://journalofbigdata.springeropen.com/articles/10.1186/s40537-021-00459-1
3. Predicting Personality From Twitter. https://ieeexplore.ieee.org/document/6113107
4. Manifestations of Personality in Online Social Networks: Self-Reported Facebook-Related Behaviors and Observable Profile Information by Samuel D. Gosling. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3180765/