Declaration of Authorship

This is to certify that thesis titled, 'Dark Triad detection and analysis from social media text' and the work presented in it is the outcome of the analysis and experiments carried out by Morsalina, Fariha Fairoz and Fariha Anjum under the supervision of Md. Mohsinul Kabir, Assistant Professor of Department of Computer Science and Engineering (CSE), Islamic University of Technology (IUT), Gazipur, Dhaka, Bangladesh. It is also declared that neither this thesis nor any part of it has been submitted anywhere else for any degree or diploma. Information derived from the published and unpublished work of others has been acknowledged in the text and a list of references is given.

Authors:

Morsalina

Morsalina
Student ID: 180041103

Fariha

Fariha Fairoz
Student ID: 180041214

Fariha Anjum

Fariha Anjum
Student ID: 180041119

Supervisor:

Md. Mohsinul Kabir

Mohsinul Kabir
Assistant Professor,
Department of Computer Science and Engineering,
Islamic University of Technology

ISLAMIC UNIVERSITY OF TECHNOLOGY

Dark Triad detection and analysis from social media text

By

Morsalina, 180041103

Fariha Fairoz, 180041214

Fariha Anjum, 180041119

A thesis submitted in partial fulfilment of the requirements for the degree of B.Sc. in Computer Science and Engineering

Academic Year: 2021-2022

Department of Computer Science and Engineering

Islamic University of Technology.

A Subsidiary Organ of the Organization of Islamic Cooperation.

Dhaka, Bangladesh.

May 2023

Declaration of Authorship

This is to certify that thesis titled, 'Dark Triad detection and analysis from social media text' and the work presented in it is the outcome of the analysis and experiments carried out by Morsalina, Fariha Fairoz and Fariha Anjum under the supervision of Md. Mohsinul Kabir, Assistant Professor of Department of Computer Science and Engineering (CSE), Islamic University of Technology (IUT), Gazipur, Dhaka, Bangladesh. It is also declared that neither this thesis nor any part of it has been submitted anywhere else for any degree or diploma. Information derived from the published and unpublished work of others has been acknowledged in the text and a list of references is given.

Authors:

Morsalina
Student ID: 180041103

Fariha Fairoz
Student ID: 180041214

Fariha Anjum
Student ID: 180041119

Supervisor:

Mohsinul Kabir
Assistant Professor,
Department of Computer Science and Engineering,
Islamic University of Technology

Abstract

The increased use of social media platform usage such as Facebook has given an opportunity to express one's thoughts and ideas to everyone. Social media posts can be used as a medium for determining different psychological traits, such as dark triad characteristics. In our thesis, we used peoples' social media posts, and using those posts we tried to detect presence of dark triad traits based on the hand crafted features extracted from their posts. We also have shown, the usage of code mixing as a effective feature. We have used traditional machine learning models, ensemble models as well as transformer based language models to find the best possible outcome. For finding linguistic features analysis we have used Polarity of text, subjectivity, lexical_density, word_tokens, word_count, avg_word_length, word_freq_dist, stopword_count, part_of_speech, Topic segmentation and many more. We then compared the performance of all the models.For all the traits, ensemble of 4 traditional machine learning models outperformed all other models.The highest accuracy achieved in narcissist detection was 96.36% , for Machiavelli it was 98.27%, and for psychopathy it was 99.62%.

Keyword - Social Media; Mental Health; dark triad; narcissist; Machiavelli; code switching; sentiment; hand crafted feature; psychopath; traits

Acknowledgements

We would like to express our heart felt gratitude to Allah Subhanu Wata'ala for giving us strength and patience to complete this study. We are indebted to our supervisor Md. Mohsinul Kabir, Assistant Professor, Department of Computer Science and Engineering, Islamic University of Technology for his meticulous mentorship. His motivation, suggestions, and insights for this thesis have been invaluable. His valuable opinion, time, and input provided throughout the thesis work helped us to do our thesis work in a proper way. would also like to express our grateful appreciation to Dr. Md. Kamrul Hasan and Dr. Hasan Mahmud for their constant motivation and support throughout this study. We are grateful to them for their constant and energetic guidance and valuable advice.

This work is supported by

Systems and Software Lab (SSL)
Department of Computer Science and Engineering (CSE)
Islamic University of Technology (IUT)

Contents

List of Figures

List of Tables

Dedicated to our parents and siblings for their lifelong dedicated support in every step of our lives

Chapter 1

Introduction

At present, with the advancement of the internet, social media platforms have become very popular for sharing ideas and thoughts of individuals. With 2.91 billion users Facebook has become the most popular online platform. People now can express their thoughts, share their photos, communicate with other people around the world, and so on using Facebook.

1.1 The context

Many people are joining Facebook every day. A huge amount of data is being generated every day through Facebook posts. These posts can be for self-expression, motivational, violence-spreading, self-promotional, anti-social posts, and many more. These posts generally express human psychology. These mischievous human traits have been categorized in the form of "dark triad" in human psychology research [8]. This dark triad is categorized into 3 traits:

1. Narcissism:
It is a self-centered personality style characterized as having an excessive interest in one's physical appearance or image and an excessive preoccupation with one's own needs, often at the expense of others.

2. Psychopathy:
Psychopathy is a collection of psychological traits and actions that are commonly connected with a lack of emotional responsiveness and compassion, recklessness, superficial charm, and insensitivity to negative consequences.

The two-factor paradigm separates psychopathy into main (Factor 1) and secondary (Factor 2) psychopathy. Primary psychopathy is characterized by interpersonal and effective elements like as coldness and callous manipulation, whereas secondary psychopathy is characterized by dangerous, impulsive activities. Primary psychopathy has also been dubbed the "successful" psychopathy, because having minimal guilt and empathy may be a fantastic strategy for gaining power in society. Secondary psychopathy, on the other hand, is the "unsuccessful" psychopathy side, leading to crime and jail rather than a career in politics.

3. Machiavellianism:
To identify the characteristic of a person by his/her online status, many research has been done so far. Research shows that people who love to update their status have a relationship with narcissism [1]. Using deep learning techniques psychopathy characteristics can be detected from social media texts [4]. Furthermore, linguistic features are also analyzed to detect human traits.

1.2 Problem Statement

Though many research have been done for detecting the dark triad of a person using online status, very few datasets are available online. Also there is no code available as far our knowledge.

Again, some papers detected only one kind of psychological trait out of the 3 dark triads [4] [2] etc. Some used psycholinguistic features to classify and analyze these traits [7]. And some of them used statistical analysis to create a dataset and applied some ML models to evaluate the performance [2].

1.3 Proposed Solution

In our thesis, we want to analyze the dark triad trait of a person using two approaches.

- 1^{st} **Approach:** We want to analyze linguistic features of posts extracted from social media through traditional machine learning models as well as their ensemble versions. And find whether there is enhancement in the precision of dark triad detection.

- 2^{nd} **Approach:** We will add code mixing/ switching information as features and determine whether inclusion of code mixing as a feature make the detection of dark triad more robust.

Chapter 2

Literature Review

This literature review is an overview of the previously published works on our research topic. We have divided our literature review into 5 parts. (i) Detection of narcissism, (ii) Detection of psychopathy (iii) Detection of Machiavellianism, (iv) Detection of dark triad, (v) Other relevant researches.

2.1 Detection of narcissism

In this section, we have discussed the kinds of literature that detect narcissism characteristics. We know, narcissism is one of the traits of the dark triad where people are obsessed with themselves.

2.1.1 Narcissistic power poster? On the relationship between narcissism and status updating activity on Facebook [1]

This paper was authored by Grosse. et al. in 2014 [1]. It was published in the journal of research in personality.

This paper studied that, people who update their status vet frequently have a narcissistic mentality. It is because, First of all, according to research done by Bergman et al. [14] in 2011, it is demonstrated that status updates may precisely cater to narcissists' ongoing need for attention and external affirmation since status updates are easily available and quickly reach a large audience that is asked to contribute feedback in the form of "likes" or comments. Second, status updates are broadcast to a broader audience rather than a single recipient. This, of course, determines the self-focus of status updates. According to a paper Bergman et al. [14], 2011, narcissists tend to be self-centered, egoistic, like to talk about themselves, and show a lack of empathy, these characteristics of status updates should suit them well.

Experiment Procedure: For conducting this experiment, they considered students from US and Germany. For students from Germany, they performed empirical research consisting of 209 students, among them 164 were female and 64 were male. Everyone got compensation of 20 euros for participating in this experiment. To assess whether narcissism predicts status-updating activity, they used Poisson regression models because their outcome – the number of posted status updates within six weeks – was a count variable with a small mean.

According to their study, they showed that people who share status frequently have narcissistic behavior.

They also conducted an online survey based on the questionnaire. This also showed the same result.

Limitations: This paper is based on an empirical study. It does not consider analyzing texts for detecting narcissism traits of human behavior.

2.1.2 Detecting Narcissist Dark Triad Psychological Traits from Twitter [2]

This paper was authored by Haz. et al. in 2022, it was published at ICAART.

This paper proposed two Machine Learning models 1) Support Vector Machine (SVM) and 2) Naive Bayes method. These two models are used to classify the comments of twitter as narcissistic or non-narsissistic traits. They have also created a manual dataset to train the models. To create this dataset they employed NLP techniques to process the comments from twitter. They have designed and applied 3 different techniques to label each of the comment.

Dataset Creation: They have created a dataset from Twitter. To collect comments from the Twitter users, they tweeted 7 tweets and these seven tweets were related to

Tweet text
¿Cómo crees que actuarías ante el sufrimiento de tu peor enemigo, podrías sentirlo como tuyo, o prefieres cambiar e ignorar el tema? (How do you think you would act in the face of the the suffering of your worst enemy, could you feel it as yours, or would you prefer to change and ignore the subject?)
¿Si a un hijo tuyo le detectan una malformación lo darías en adopción, o preferirías que muera para no tener que cuidarlo o que no sufra discriminación? (If your child is found to have a malformation, would you give him/her up for adoption, or would you prefer him/her to die so you do not have to take care of him/her, or he/she does not suffer discrimination?)

Table 2.1: Example of posted triggering tweets

the main traits of narcissism characteristics. Before posting these tweets, they took validation of these tweets from two psychologists. Examples of tweets are listed in table 2.1.

To ensure that the number of participants is large, they shared those tweets on Facebook, mail, and twitter. People were asked to fill out a form that contained NPI test. NPI is a test that measures narcissism in the normal population. The version of NPI that they used has 40 questions. They collected comments for 6 weeks and they filtered out some of the comments.

They applied three methods to analyze the dataset.
1. They at first considered the NPI online test. And they labeled the tweets as narcissistic if the user has a narcissistic score on NPI test.
2. They used a dictionary of words that consists of words that are frequently used by narcissists. They established the two-word rule, which checked each word of an obtained comment with the words stored in the dictionary. Then, if at least two words were found in the dictionary, the comment was labeled as narcissistic, otherwise non-narcissistic.
3. They used a manual approach which was conducted by two psychologists. The process is that, if two psychologists classified a tweet as narcissistic then the tweet was labeled as narcissistic.
Table 2.2 shows the number of narcissistic tweets.

Classifier: The full process is divided into two parts 1) pre-processing the data where the text will be represented by Vector Space model and 2) a supervised model is developed to classify a tweet as narcissistic or non-narcissistic.

Method	**Narcissistic**	**Non-Narcissistic**	**Total**
NPI Test	93	312	405
Dictionary	119	358	477
Manual Evaluation	98	112	210

Table 2.2: Number of narcissistic posts

For pre-processing data, they first removed punctuations, emojis and stop words. Then the tokenized the sentence into a set of words. These words were filtered by Bag of Words model. This model was used to extract some specific terms.

The basic pre processing is followed by the upcoming steps, such as passing the values to the models. Two of the most used such traditional models are support vector machine and naive bayes. It calculates the posterior probability considering there is no inter dependency among the values. After that they built two ML models such as SVM and naïve bayes to classify the sentences.

The architecture of the proposed system is shown in figure 2.1

Finally, they trained the models using the dataset. And then tested it. The result was quite good in this case. For Naïve Bayes classifier, the test accuracy was 79%, whereas it was 74% for SVM. In the case of the manually created dataset, the accuracy was 74% for both of the models. The models performed comparatively good in the case of the dictionary based.

Limitations: For text processing, this paper considers only individual words, it does not consider the context of the words. Again this paper only considers about narcissism.

2.1.3 How did you like 2017? Detection of language markers of depression and narcissism in personal narratives [3]

This paper is authored by Rathner et. al. It was published at ISCA in 2018.

This paper focuses on the word use in case of depression and narcissism i,e. word usage of depressed or narcissist people using LIWC(word count).

The study was conducted on 220 individuals of age greater or equal to 18. There were all type of people married, unmarried, student, job-holders, unemployed etc. But the

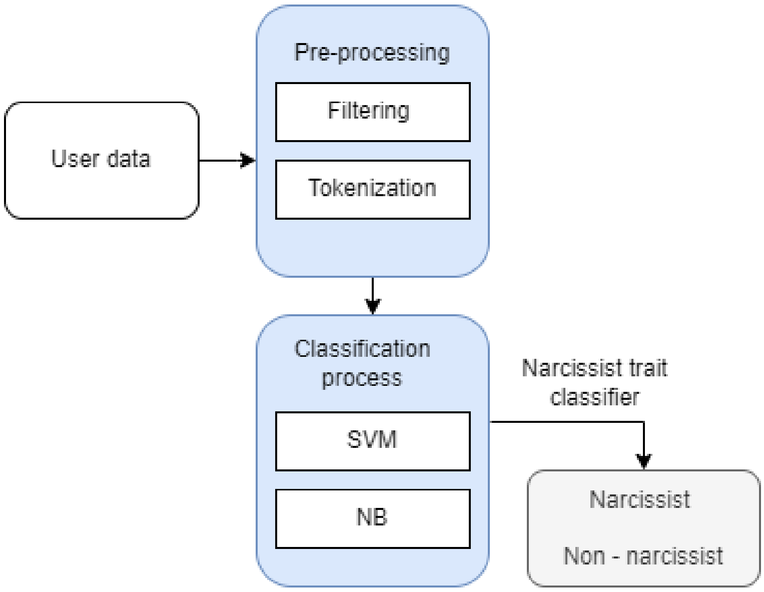


FIGURE 2.1: System Architecture

only condition was that they all had to have German as mother tongue to avoid effect on language use. Among them, 200 were never diagnosed with mental instability, 16 were highly depressed and 4 had other problems.

How this experiment was conducted: The participants were instructed to report their age, marital status, employment status, mother tongue and mental health diagnosis. Next they were told to write an essay on their life plans for the ongoing and upcoming year. At last they were instructed to fillup the NPI-d ans BDI-II questionnaires. BDI-II is the test for presence and severity level of depression in people and NPI-d is the test for narcissiom scale in people. BDI-II has 21 items rated on 4 point-scale and NPI-d has 17 items on 5 point-scale. The standard deviation of BDI-II and NPI-d were found to be 9.13 and 10,74 respectively. Next the essays were analysed using LIWC of German dictionary. Lastly linear regression was performed on the word categories. Gradually unimportant predictors got excluded giving a precise result.

Result: The result has two parts - one on depression and another on narcissism. In case of prediction on depressive symptoms depending on words use, use of negative words and anxiety related words are deemed to be highly possible symptoms for depressed people. Well, the words rate may vary from person to person depending on their level of mental instability. And in case of narcissism, we, affect and negative emotional words are predicted to be highly likely to be used by narcissists.

Limitations: One of the major limitation of this work is the exclusion of prediction categories as the excluded predictor may have significant influence but ignored. Still this approach was taken to avoid irrelevant predictors and huge iterations due to them. Another limitation is in the age restriction. Nowadays, a lot of teenagers are facing huge pressure academically or in personal life and suffering from huge depression. So to say, a huge group of people have been excluded from the research.

2.2 Detection of psychopathy

In this section, we have discussed the kinds of literature that detect psychopathy characteristics. We know, psychopathy is one of the traits of the dark triad where people do anti-social behavior.

2.2.1 Detection and Classification of Psychopathic Personality Trait from Social Media Text Using Deep Learning Model [4]

The goal of this paper is to classify the text into psychopaths and non psychopath classes using a deep learning model, namely, BiLSTM.

maximum accuracy of 85% and performs better as compared to other Bi-LSTM models with parameters (batch size = 32, Bi LSTM unit = 250, and vocabulary size = 2000).

preprocessing is done as follows: Case Transformation. Elimination of Special Characters. keras tokenizer keras embedding layer. drop out layer BiLSTM Layer

Limitations:
(1) The dataset collected for experimentation is insufficient that may decay the performance of the proposed model (2) The present study is limited to the implementation of the BI-LSTM model without applying the fusion of different deep neural networks such as CNN + LSTM, CNN + Bi-LSTM, CNN + RNN, and CNN + Bi-RNN (3) They exploited only the random word embedding for the input layer, while other different word representation techniques like Glove and FastText may improve the system performance

2.2.2 A Hybrid CNN-LSTM Model for Psychopathic Class Detection from Tweeter Users [5]

The main task of the paper is to classifying the text into psychopaths and non psychopath classes using deep learning model, namely, CNN + LSTM.

the proposed CNN-LSTM model operated in the following four layers: (i) the embedding layer, which acted as the input layer; (ii) the CNN, which was used as the hidden layer; (iii) the LSTM, which also worked as the hidden layer; and finally, (iv) the dense layer with the SoftMax activation function, which acted as the output layer. It was observed that the CNN-LSTM model with LSTM unit size 140, filter no. 16, and wit a filter size of value 5, achieved a maximum accuracy of 91%. It was also observed that the performance of the model was improved gradually by reducing the values of the following parameters: LSTM unit size, filter size, and filter number.

The dataset was created by twitter crawling tweets with #psychopath tag. Later, manual annotation was done by assigning the task to three human annotators, who are professional psychiatrists. Each of them assigned each tweet a class label: "psychopath" or "non-psychopath". In this way, they received three votes for each tweet. The class label is selected on the basis of majority voting scheme. For basic pre processing, they did Case Transformation and elimination of Special Characters.

Limitations:
dataset size is small, there were only 601 entries.Effect of pretrained embedding is not yet testified.The current study used random word embeddings. So it is expected to have better performance with pre trained embeddings. LSTM unit was within 140.The results are not observed for more than 140 units.

2.3 Detection of Machiavellianism

In this section, we have discussed the kinds of literature that detect machiavellianism characteristics. We know, machiavellianism is one of the traits of the dark triad where people try to manipulate others for their own interest.

We did not find any paper which solely worked on machiavelli detection.

2.4 Detection of dark triad

In this section, we have discussed about the papers that detect all of the traits of dark triad altogether.

2.4.1 Predicting the Dark Triad for Social Network Users using Their Personality Characteristics [6]

The main task of the paper is to classify the text into Dark triad classes using the regression technique.

the proposed method initially did feature extraction using IBM Watson to determine 17 features. Later those are divided into 3 sets, set 1 with 5 features, set 2 with 12 features, and set 3 with 17 features. The highest average trait accuracy was found in set 3 with linear regression which is 70.3%

Dataset Preparation: In the data collection module, Twitter handles (i.e., users' IDs) are used by a Twitter API for the purpose of crawling users' posted tweets. All extracted tweets for each user are then concatenated in a single document.

Feature extraction: IBM Watson Personality Insights API is a cloud-based service that uses machine learning to analyze the emotional content behind unstructured customer communications like emails, social posts, and blogs. Here the researchers used this API to extract features.The extracted Values features are based on Schwartz's personal values theory and Maslow's hierarchy of needs.In total there are 17 features extracted and three sets of dataset are made taking 5, 12 and 17 features in each set.

Methodology: The dataset is prepared with 863 users which later came down to 564 after revision.The results found in the dirty dozen questionnaire is used to produce two thresholds, one is done using median split and another is discriminative values and scaled to fit log based class values.The algorithms used are Linear Regression with elastic net regularizer, Normal Linear Regression, Logistic Regression (LR) and random Forest. These are used for prediction to evaluate the proposed category features sets. Above mentions algorithms are particularly well known for dark triad detection.

Result: For the two thresholds, the Linear Regression with an Elastic Net regularizer performs better than the Normal Linear Regression. For both the trait and average

Table 1: Schwartz's personal values

(A)	Personal Values	Description
a.1	Self-transcendence	Overcome the limits of self and show concern for the interests of others.
a.2	Conservation	Emphasize resistance to change and self-restriction.
a.3	Hedonism	Seek pleasure and sensual enjoyment.
a.4	Self-enhancement	Seek personal success.
a.5	Openness to change	Emphasize independent action, and willingness for new experiences.

Classification Threshold	Classifiers	Features	TRAITS' classification Accuracy (%)			Avg. Trait Accuracy (%)	Avg. Classifier Accuracy (%)
			Machiavellianism	Narcissism	Psychopathy		
(θ1)	LR	Set (I)	73.4	61.0	75.0	**69.8**	**69.9**
		Set (II)	72.2	61.5	75.0	**69.6**	
		Set (III)	73.8	61.5	75.7	**70.3**	
	RF	Set (I)	64.5	52.1	78.1	**64.9**	**66.8**
		Set (II)	69.2	58.6	78.7	**68.8**	
		Set (III)	65.7	54.4	79.9	**66.7**	
(θ2)	LR	Set (I)	69.8	69.2	62.7	**67.3**	**66.6**
		Set (II)	69.0	66.9	63.3	**66.4**	
		Set (III)	69.2	68.6	60.9	**66.3**	
	RF	Set (I)	65.0	63.9	51.5	**60.1**	**62.4**
		Set (II)	68.6	62.7	59.7	**63.7**	
		Set (III)	67.5	62.7	59.8	**63.3**	

accuracies, threshold 1 is somewhat superior to threshold 2. Similar to the RMSE, Set (III) produced the best forecast with an average accuracy of 69% at the 1 level. The results indicate that the combined Values and Needs feature set (III) is still the best for trait and average accuracies at 2. In this instance, the average accuracy is 68.3%. The Narcissism prediction improves by 13.4% when using the reference-based threshold (2), indicating the high predictive ability for this trait, when comparing the results of both classification thresholds on each trait level..

2.4.2 Studying the Dark Triad of Personality through Twitter Behavior [7]

This paper was authored by Daniel Preo¸tiuc-Pietro, Jordan Carpenter, Salvatore Giorgi, Lyle Ungar. This paper was published in CIKM in 2016.

The paper used language features to detect the dark triad personality of a person. They employed models from natural language processing (NLP) and image analysis in order to extract features that are both predictive and interpretable, such as word topics or

Machiavellianism
I tend to manipulate others to get my way
I have used deceit or lied to get my way
I tend to exploit others towards my own end
I have used flattery to get my way
Narcissism
I tend to want others to admire me
I tend to want others to pay attention to me
I tend to seek prestige or status
I tend to expect special favors from others
Psychopathy
I tend to lack remorse
I tend to be callous or insensitive
I tend to be unconcerned with the morality of my actions
I tend to be cynical

Table 2.3: Dirty Dozen Questionaire

facial features. They used existing state-of-the-art models that predict the expression of emotion and relate them to previous findings. Secondly, they aimed to build a predictive model for the dark triad traits that use only public Twitter information.

As there is no dataset available on dark triad, so built a dataset. The process is described below:

Dataset Preparation: They have collected data from Twitter users who have taken twelve questionnaires from "dirty dozen" and given their public and real Twitter handles. The dirty dozen questions are listed in table 2.3. This uses four questions to assess each of the three traits. These three traits are psychopathy, Machiavelli, and narcissism. The trait score is the arithmetic mean of its four questions scored with values on a 1–5 scale. The questions are presented in Table 1. Additionally, we use a combined dark triad score as the arithmetic mean of all twelve questions, Similar to previous research to detect dark triad traits.

At first, the dataset contained 538,712 posts from 710 distinct users. But to gain confidence in text analysis, they considered tweets that have a minimum of 500 tokens in length, which made the size of the dataset to 536,579.

They used a feature extraction method for text analysis. For each of the users, they considered the following feature extraction methods:

- They have represented the posts using a normalized frequency distribution over a vocabulary. That is, they have used "bag of words" representation. They have used "unigram" in this case, each word is a feature.

- Secondly, they have used LIWC to represent the texts of the post. automatically counts word frequencies for 64 different categories manually constructed based on psychological theory. Each user is thereby represented as a frequency distribution over these categories. That is how many times user x, has used the specific category.

- Then they used "Word clustering" to reduce the feature dimension. Similar words are in the same cluster. To create these groups of words, they used an automatic method that leverages word co-occurrence patterns in large corpora by making use of the distributional hypothesis: similar words tend to co-occur in similar contexts. They used a separate reference corpus of almost 400 million tweets to compute a word-to-word similarity matrix using Word2Vec. This is a neural network approach to distributed word representations, where the words are projected into a lower dimensional dense vector space. In this low-dimensional vector space, words with a small Euclidean distance are considered semantically similar. They then applied spectral clustering to obtain hard clusters of words from the word $\times$ word similarity matrix. According to them, a cluster size of 200 works better than that of 100, 500.

- Finally they have predicted and assigned emotions with these texts. They have used a model name "EKMAN model" which has 6 emotions: 1. Anger, 2. Fear, 3. Joy, 4. Sadness, 5. Disgust, 6. Surprise. They used the publicly available crowd sourcing-derived lexicon of words associated with any of the six emotions, as well as trust and anticipation and general positive and negative sentiment. They they have assigned emotion with each of the messages and took the average across all users to obtain user-level emotion.

They have also extracted features of images and profile information for their dark triad analysis.

To analyze their result, they explored the relationships between Twitter behaviors and the three dark triad traits plus the overall dark triad score, see image 2.2. They used univariate Pearson correlations between each feature and the logarithm of the dark triad

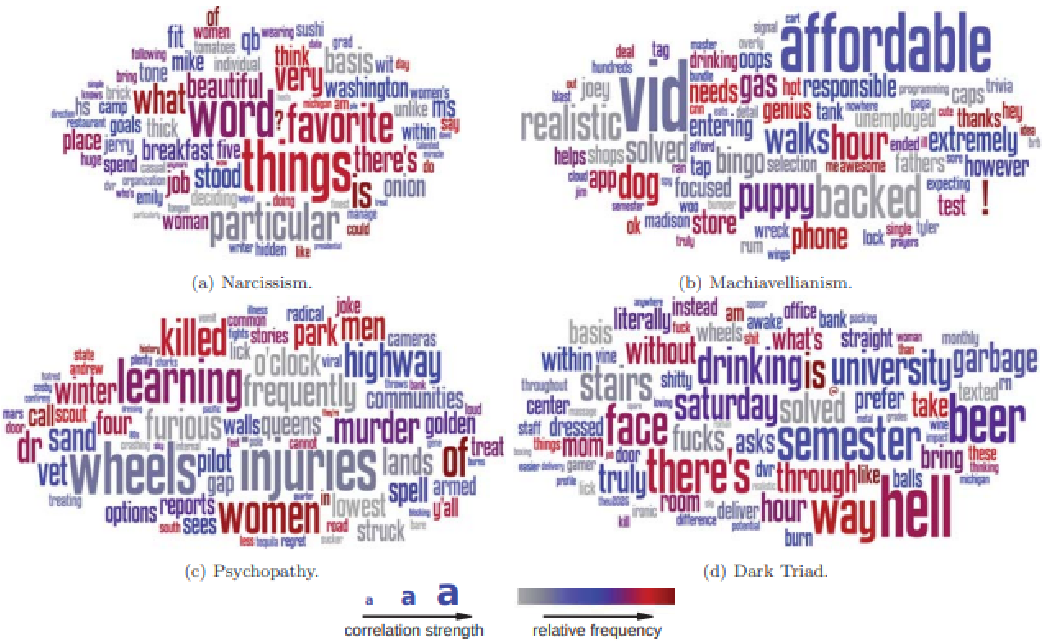


FIGURE 2.2: The word clouds show the unigrams with the highest Pearson correlation to each of the dark triad traits

scores. Their findings are listed here.

Finally, they used all the previously derived features to create a predictive model of the three dark triad traits, as well as a composite score. They considered this a regression problem to which they can apply machine learning algorithms. They used a linear regression algorithm with an Elastic Net regularizer with the ScikitLearn implementation. They observed that, with the exception of psychopathy, the textual topic features obtain the best prediction results. In the case of psychopathy, the LIWC textual categories obtain a better performance.

Limitations: The population using social media, here Twitter, and in addition Amazon Mechanical Turk is not representative of the general population.

2.4.3 The Dark Triad of personality: Narcissism, Machiavellianism, and psychopathy [8]

This paper was authored by Delroy L. Paulhus and Kevin M. Williams. It was published in 2002 in the journal of Elsevier.

This paper focuses on the feature extraction of the students of the three traits - narcissism, psochopathy and machiavellianism and their correlation.

How this experiment was conducted: This experiment was conducted on 245 students. They were given some questionnaires to answer regarding these three traits. They submitted those answers which were then analysed.

Result: Males scored significantly higher on all three of the Dark Triad. The correlation was seen among almost all the three traits of dark triad. The correlation analysis found : (machiavellianism, psochopathy) - .31, (psochopathy, narcissism) - .50, (machiavellianism, narcissism) - .25. Subclinical psychopaths were distinguished by low neuroticism; Machiavellians, and psychopaths were low in conscientiousness; narcissism showed small positive associations with cognitive ability. Narcissists and, to a lesser extent, psychopaths exhibited self-enhancement on two objectively scored indexes.

Limitations: This paper mainly focused on the correlation among the three traits of dark triad. The authors also said that there about some feature scales in these three traits. But they did not say anything about the actual features of the traits i.e. which features are more seen in which traits or which features make them say that the student is narcissist, machiavellian or psychopath.

2.5 Use of LIWC tool in personality detection

LIWC is a psycholinguistic feature extraction process.It is used in many researches for extracting features.

2.5.1 Predicting Personality from Social Media Text[9]

This paper replicates text-based Big Five personality score predictions generated by the Receptiviti API—a tool built on and tied to LIWC. The LIWC tool is a widely tested, validated, and applied system for performing psycholinguistic text analysis. In 2015, Receptiviti was released along with new version of LIWC. Receptiviti includes some added feature analysis than LIWC. This paper focuses on personality attributes in Receptiviti which provides estimates of Big Five personality traits for a user based on a text sample from dataset.

Dataset Preparation: The Receptiviti API allows users to submit a text sample to be evaluated, and a graphical depiction of some qualities as well as a JSON-based result list is provided. To evaluate the accuracy of the API's Big Five personality estimations, they provided social media-derived text from persons with known Big Five scores. They used social media data sets acquired in prior personality study studies, two from Facebook and one from Twitter, as well as additional Facebook data sets from the myPersonality project. These data sets include 10,000 status updates for 250 people with personality ratings, as well as a big database of 22 million status updates for 154,000 users with personality scores for each user. They combined the status update and personality databases to produce the myPersonality Large data set for their investigation. The data sets were formed of text contributed by users on the social media network and combined into a single document. Because the Receptiviti API requires at least 300 words for statistical correctness, they eliminated all individuals with fewer than 300 words. They had data from roughly 9,000 social media users and limited the word count to 10,000 for individuals who had contributed a lot of text. They then submitted the user-representing document to the Receptiviti API and collected the raw Big Five information for each user. They normalized the Receptiviti scores and adjusted the scores to the 1-5 scale because the personality values from Receptiviti were not on the same 1-5 scale as their input data2. This allowed for a direct comparison of known and expected values for each subject.

Methodology: They calculated the Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Pearson Correlation Coefficient between the known and projected values for each of the five characteristics in each data set. Table 3 displays the outcomes. With an MAE more than 1.0, openness was the most prone to mistake. Conscientiousness and Neuroticism had the greatest prediction performance, with an MAE of less than 20%. Correlations between expected and known values for a specific personality characteristic ranged from weak to near-zero. They tested this by comparing performance of users with more than 2,000 words versus those with fewer than 2,000 words. They discovered no significant differences on any characteristic in any of the three smaller data sets (p0.05). This is confirmed further by the fact that the huge myPersonality data collection, with greater average words per individual, performed no better than the smaller data sets.

Result: The Receptiviti API performed similarly on each personality characteristic across all four data sets. MAE-based error rates were frequently in the 15-30

Limitations: They compiled social media messages into ”documents” of at least 300 words. When a single document or social media post of 300 words or more is evaluated, Receptiviti may offer more accurate findings. However, they did not have enough long social media posts to do this sort of analysis.

2.6 Feature extraction techniques

Though deep learning models are leading in almost all kind of pattern recognition task, there are still some fields where specific features are need to be checked to draw a conclusion. Such features are generally selected by specialists.

2.6.1 Sentiment Analysis of Twitter Data: A Survey of Techniques[10]

This study includes a survey and comparative analysis of existing methodologies for opinion mining, such as machine learning and lexicon-based approaches, as well as assessment criteria. It also examines the problems and uses of Sentiment Analysis on Twitter, such as unstructured, diverse viewpoints and the usage of machine learning methods such as Naive Bayes, Max Entropy, and Support Vector Machine. It also examines the problems and uses of sentiment analysis on Twitter, such as unstructured, diverse viewpoints and the usage of machine learning methods such as Naive Bayes, Max Entropy, and Support Vector Machine.

The Internet has revolutionized the way individuals express their thoughts, ideas, and exchange information about their everyday lives. Twitter, status updates, blog posts, comments, reviews, and other forms of social media generate a significant volume of sentiment-rich data. Businesses utilize this data to interact with their consumers for advertising purposes. Before purchasing a product, customers can utilize sentiment analysis (SA) to determine whether the information provided is adequate. Textual information retrieval strategies are concerned with processing, searching, or evaluating the factual material that is available. Because of the massive expansion of available information on online sources such as blogs and social networks, SA provides numerous challenging chances for developing new applications. For example, suggestions of things offered by a recommendation system can be anticipated by taking into account factors such as favorable or negative comments about such products.

Dataset Preprocessing: The Twitter dataset utilized in this survey work is already classified into two classifications - negative and positive polarity - making sentiment analysis of the data simple to detect the influence of various characteristics. Raw data with polarity is prone to inconsistency and redundancy. The tweet was preprocessed to take into account factors such as URLs, spelling, symbols, and so on.

Feature extraction: The feature extraction approach pulls characteristics from the preprocessed dataset that are utilized to compute the positive and negative polarity in a phrase. Machine learning approaches need expressing essential properties of text or documents as feature vectors, which are then utilized for classification tasks. The following are some examples of characteristics that have been documented in the literature:

- Words And Their Frequencies
- Parts Of Speech Tags
- Opinion Words And Phrases
- Position Of Terms
- Negation
- Syntax

Methodology: They trained the classifier via Naive Bayes, Max Entropy, and Support Vector Machine. There are primarily two ways for sentiment analysis in Twitter data: machine learning-based and lexicon-based. Machine learning involves both supervised and unsupervised training. Lexicon is built on dictionaries and corpora. They performed sentiment analysis at the word, sentence, document, and feature levels.

Result: Machine learning approaches, such as SVM and naive Bayes, have the best accuracy and may be considered the baseline learning methods, whereas lexicon-based methods are extremely successful in some circumstances, requiring little work in human-labeled documents.

2.7 Use of deep learning models in psychology

Deep learning methods help in studying personality, behavior, emotions, and sentiments. IT is now widely used for prediction and pattern analysis in psychological researches.

2.7.1 Recent trends in deep learning based personality detection[11]

This study examines key machine learning models for personality recognition, with a focus on deep learning-based approaches. It includes an overview of the most prevalent techniques to automated personality recognition, as well as diverse computational datasets, industrial applications, and cutting-edge machine learning models. Personality identification is a large and complex field, and this study focuses solely on computational techniques.

Personality is a collection of an individual's behavior, emotions, motivation, and mental processes. It has a significant influence on their lives, influencing their life choices, well-being, health, and preferences and wants. The Woodworth Psychoneurotic Inventory (Papurt 1930) is widely regarded as the earliest personality test, having been devised during World War I for the United States military to screen recruits for shell shock risks. Taibi Kahler created the Process Communication Model (PCM) with NASA support, and it is currently mostly used for consulting purposes to assist individuals become more effective communicators and avoid or resolve problems when communication has gone off track. Within the first 10 seconds of meeting someone, people generate an impression about them based on their facial traits, and these perceptions impact their conduct toward a newly encountered person or a human-like robot.

Various approaches for personality modeling have been employed by researchers. The Myers-Briggs Type Indicator (MBTI) (Briggs-Myers 1993) is one of the most frequently used personality tests in the world, being administered to millions of employees in thousands of businesses each year. The MBTI personality test divides people into two categories in each dimension: introversion against extraversion, sensing versus intuiting, thinking versus emotion, and judging versus perceiving.

In today's world, computerized personality identification systems have a wide range of industrial uses. In this subject, research is increasing, resulting in improved models with more accuracy and dependability. Artificial personality can be embedded into practically all human-computer interactions, allowing for more intriguing and individualized encounters. Personality traits can also be employed as inputs for other tasks such as sarcasm detection, deception detection, and word polarity disambiguation systems.

Various methods for accessing personality have been created and perfected during the last century. However, there are ethical concerns about using these exams inappropriately in terms of invasion of privacy, cultural prejudice, and confidentiality. Political parties have recently begun to experiment with large-scale machine learning-based personality recognition for political forecasting. The Cambridge Analytica data scandal, which entailed the acquisition of personally identifiable information from 87 million Facebook users, is one example of this. Algorithmic Impact Assessments (AIAs) are proposed by the AI Now Institute to address biases in machine learning systems.

DARPA proposes Explainable Artificial Intelligence (XAI) as a solution. AIAs strive to give clarity to the public by publicly disclosing and describing algorithm systems utilized while allowing the public to dispute these systems, providing audit and evaluation mechanisms, and strengthening the internal capacity of public agencies to comprehend the systems they employ.

Summary of the various popular models, architectures and techniques commonly used in deep learning-based personality detection:
Text – feature extracted by LIWC, MRC. Models used are SVM, Naïve bayes, LSTM etc.
Audio - feature extracted by MFCC, ZCR etc.
Visual – facial feature by VGG(CNN)
Many other models are used for bimodal or trimodal features.

The necessity for larger, more accurate, and diversified datasets for automated personality recognition is discussed in this research. Almost all existing datasets are centered on the Big-Five personality model, with relatively few focusing on additional personality measures such as the MBTI or PEN. Recent multimodal deep learning approaches have performed well and are beginning to predict personality. Deep models have become the new cutting-edge approaches not only for personality recognition, but also in other domains. They anticipate that this trend will continue with deeper models and new architectures capable of mapping extremely complicated functionalities.

2.8 Use of Subjectivity and polarity classification as a feature

Subjectivity tells how a person's point of view results in his perception i.e. how his perception emerges from his personal feelings or experiences. Polarity classification is a range of score that depicts the sentiment level of a particular text or message. In this section we have discussed the researches using these two as features.

2.8.1 A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts[12]

This study focuses on sentiment analysis, which attempts to detect the viewpoint behind a text span. The authors utilized a unique machine-learning strategy that applies text-categorization algorithms just to the subjective parts of the document.

Because of the possible applications, the computational treatment of opinion, emotions, and subjectivity has recently received a lot of interest. Previous techniques concentrated on picking suggestive lexical characteristics and categorizing a document based on the quantity of such features found anywhere within it. They suggest classifying phrases in the document as subjective or objective, deleting the latter, and applying a normal machine-learning classifier to the extracted data. Subjectivity extracts properly replicate the sentiment information of the original materials in a considerably more compact manner, according to their findings. They investigate extraction approaches based on a minimal cut formulation, which offers an efficient, intuitive, and successful way of combining inter-sentence level contextual information with classic bag-of-words characteristics.

Methodology: The authors employed document-level polarity classification, which is a subset of text categorization using sentiment-based categories. Standard machine learning classification approaches, such as SVMs, may be applied to the complete set of texts. However, deleting objective statements can enhance polarity categorization. To that purpose, they presented a subjectivity detector to identify whether or not each statement is subjective. Subjectivity identification may be accomplished by applying a normal classification algorithm to individual phrases, but modeling proximity connections between sentences would allow for coherence. As a result, they developed algorithms using pair-wise interaction information, such as indicating that two phrases should have the same subjectivity label. They then used cut-based categorization.

Subjectivity Detection: Their tests involve categorizing movie reviews as favorable or negative, which is an interesting assignment for a variety of reasons. On the dataset of reviews, they used support vector machines and Naive Bayes. The subjectivity detector was then trained using more reviews. They adopted the minimum-cut paradigm for subjectivity extraction, which includes background information. They discovered that while contextual restrictions are easily included using the minimum-cut formalism, they are not natural inputs for ordinary Naive Bayes and SVMs.

Result: Subjectivity detection can condense reviews into shorter excerpts while retaining polarity information. Using the minimum-cut architecture leads in efficient sentiment analysis algorithms, which can lead to statistically significant improvements in polarity-classification accuracy.

2.9 Other relevant research

In this section, we have discussed the papers that are related to personality detection and other relevant works.

2.9.1 Predicting Personality Using Facebook Status Based on Semi-supervised Learning [13]

This paper detects big five personality traits. This paper is authored by Heci Zheng and Chunhua Wu. It was published in 2019 at ACM conference.

In this paper, they introduced a semi supervised learning to analyze personality based on unlabeled data. Besides, for making full use of the language information in social media status, they adopted n-gram model to extract linguistic features. The experimental results demonstrate the semi-supervised learning can take advantage of unlabeled data and improve the accuracy of prediction model.

The contributions of this paper are:
1. They have built a prediction model using a semi-supervised learning algorithm. This algo can use unlabeled data to improve the accuracy of prediction model

2. They have used uni-gram model to extract the features of the language.

3. They found many new meaningful links between single word use with personality traits.

FIGURE 2.3: Three simple graphs

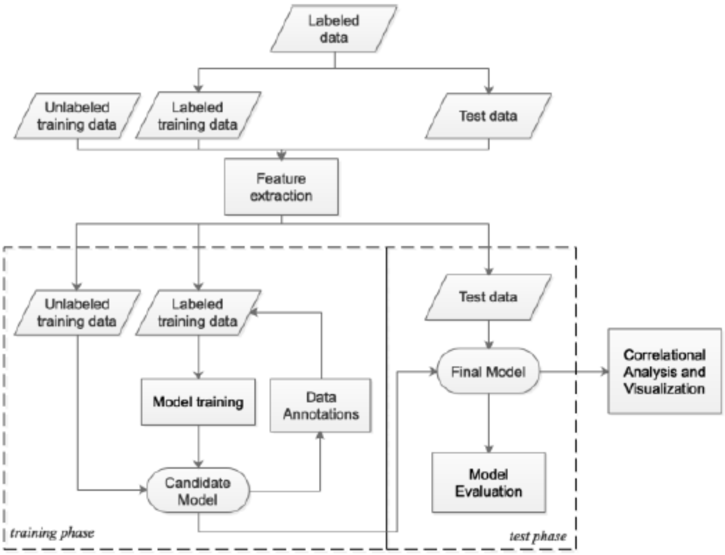


FIGURE 2.4: Model Architecture

Proposed Approach:

Their proposed model architecture is shown in the figure-2.4.

They have used a semi-supervised algorithm “Pseudo Multi-view Co-training (PMC)” to create the personality prediction model. To extract adequate linguistic features, they used both LIWC and n-gram, and also used a method based on word embedding to avoid the sparsity of word-based features. The PMC algorithm is given in the following table2.2:

For training the model: First, they downloaded the public dataset in myPersonality project, and part of the data is reserved as the test data for model evaluation. The remaining data are used as labeled training data. At the same time, they crawled a large amount of Facebook status as unlabeled training data. Then, they extracted the linguistic features of all data, and use a semi-supervised learning algorithm to train the model on labeled training data and unlabeled training data. Finally, they used test data to evaluate the performance of the model.

Lastly, the word clouds show the relationship between single words and personality. The following pictures 2.3 depict the relationship.

Chapter 3

Challenges

These are some challenges that we faced so far.

- As we have proposed to use pre trained model in our initial proposal. But we failed to get the code repository in spite of trying to reach out the corresponding authors of the papers.
- Also, each person has 100 posts, where some are lost due to it being a shared content, causing the number to decrease and number of posts became uneven, which made it harder to compare sentiment graph among the users.
- We also attempted to work on feature extraction mechanisms but there are already extensive work on dark triad and related feature extraction methodology.
- The number of the posts is also a concern.Because to predict the actual condition of a personality, this number of posts and the timeline may be inadequate to reflect the original scenario.

Chapter 4

Methodology

Our thesis investigates the detection of the dark triad personality traits utilizing hand-crafted features from social media posts, as well as the impact of code-switching as a feature. The three traits which combinely known as dark triad, namely Machiavellianism, narcissism, and psychopathy are known to have major detrimental effects on individuals and society. As a result, detecting these characteristics is critical for identifying potentially dangerous persons and preventing undesirable effects. It can also be used as a primary screening tool for identifying the possibilities of future deviations.

Existing research on dark triad detection using machine learning techniques has mostly focused on employing generic text-based characteristics, which may not adequately capture the intricacies of a person's psychological traits. Individuals may be mis classified as a result of this limitation, which can have serious consequences.

Being a sensitive research topic in psychology, there are many observations from the researchers regarding which features have more impact on determining traits than others.Some psychological behaviors have explicit reflection in their expression in terms of writing and speaking.

Social media has become a mirror of our personality, as it reflects our thoughts, emotions, and behavior. The content we post, like, and share on social media platforms provides a glimpse into our personality traits, such as extraversion, openness, and neuroticism. Research has shown that social media data can be used to accurately predict individuals' personality traits.

In a study by Kosinski et al. (2013), the researchers analyzed Facebook likes of users and found that certain likes were associated with specific personality traits. For instance, liking pages related to fashion and shopping were associated with high levels of openness, while liking pages related to sports and cars were associated with low levels of openness. Similarly, liking pages related to video games and computers were associated with high levels of introversion.

In another study by Youyou et al. (2015), the researchers analyzed the language used by users in their Facebook status updates and found that certain words and phrases were associated with specific personality traits. For example, using words related to emotions and social relationships was associated with high levels of extraversion, while using words related to analytical thinking was associated with high levels of openness.

These studies demonstrate how social media has become a mirror of our personality, as the content we post and the pages we like reflect our personality traits. Social media data can be used to gain insights into individuals' personality traits, which can have important implications for targeted advertising, job recruitment, and mental health assessments.

The methodology fig 4.1:

4.0.1

Deep learning algorithms are widely applied in many different disciplines, according to recent advances. However, it is crucial to pinpoint certain traits that set apart diverse personality types in psychology-related study fields. Additionally, it's crucial to protect participant anonymity while gathering personality type data utilizing participant postings. Because deep learning models like BERT need access to a lot of participant data, their usage might jeopardize privacy. In order to analyze personality types, this study used hand-crafted characteristics that were carefully chosen based on their applicability to personality analysis. The researchers searched for particular characteristics, including language complexity, sentence length, and word frequency. Additionally, they looked at how first-person pronouns are used and whether they are used in a forceful or tentative manner as markers of personality characteristics.. We then compared the performance of the hand-crafted features with that of commonly used deep learning models such as multilingual BERT. The results showed that while multilingual BERT effective,the hand-crafted features demonstrated superior performance. The use of hand-crafted features remains a viable alternative in situations where participant privacy is a concern. The study highlights the importance of identifying specific linguistic features that can be used to accurately predict personality types. The use of hand-crafted features is an effective approach in situations where participant privacy is a concern, but deep learning models such as BERT remain an effective tool for analyzing text data. Further research

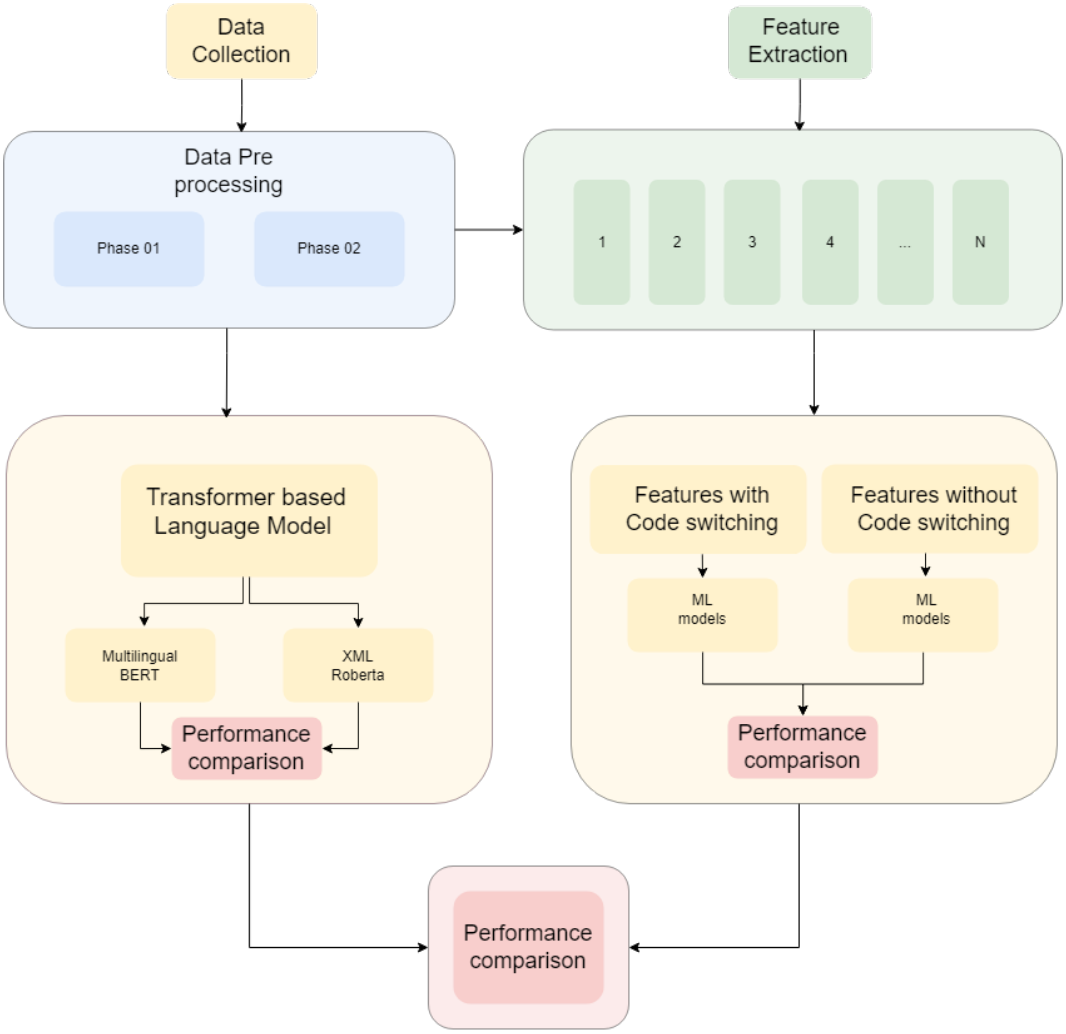


FIGURE 4.1: Methodology

is needed to identify more specific linguistic features that can accurately predict personality types and ensure participant privacy.We also analysed the impact of code switching data on analysing the dar triad personality.

4.1 Dark Triad Dataset

Here we discussed about how we have collected and preprocessed our data.

4.1.1 Data Collection

In this section we discussed how we collected data, preprocessed them and created the dataset. Refer to figure 4.2 for a better understanding.

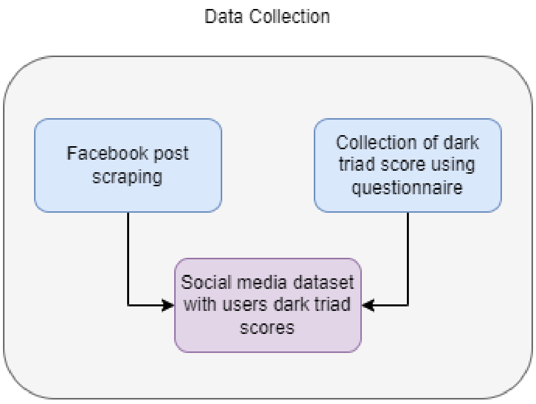


FIGURE 4.2: Data Collection

Our process includes directly obtaining the user's Facebook status with their prior consent and providing the result found through established questionnaire as label . We conducted three exams for each dark triad personality to determine our class designations.

collecting information through social media scrapping is not new in quantitative research field. Specially such initiatives are previously seen in many datasets. Social media scraping is a common method in psychological research to gather large amounts of information about human behavior. This data can be used to study a variety of topics, including mental health, social interaction, and political attitudes.[15] For example, researchers have studied the spread of misinformation through social media, the effects of social media on self-esteem, and the relationship between social media and depression.[16]

	Unnamed: 0	Mach	LSRP12	LSRP2	NRSM	english_text
0	0	68.0	1.3	2.5	3.0	my heart my soul..T-T
1	1	68.0	1.3	2.5	3.0	Free speech is a joke in this country
2	2	68.0	1.3	2.5	3.0	Make 500 sequels I don't care.Make another Jus...
3	3	68.0	1.3	2.5	3.0	Everybody wants a piece of the politically cor...
4	4	68.0	1.3	2.5	3.0	And that's how a movement dies.

FIGURE 4.3: Raw Dataset

4.1.1.1 Labelling

While **Openpsycometrics.org** offers widely recognized standard tests for each of these three, we created our own HTML page for reasons of privacy, depending on the principles these standard tests are based on (such as how many points are awarded for each question).There was three separate test for psychopath, narcissism and Machiavelli.

topic_distributions	polarity	subjectivity	flesch_reading_ease	...	RBR	EX	UH	FW	RBS	WP$	eng_usage_ratio	bng_usage_ratio	eng_switching_ratio	bng_switching_ratio
[(0, 0.03334016), (1, 0.033336256), (2, 0.0333...	0.0	0.0	118.18	...	0.0	0.0	0.0	0.0	0.0	0.0	100.000000	0.000000	0.000000	0.000000
[(0, 0.41998866), (1, 0.020005966), (2, 0.0200...	0.4	0.8	105.66	...	0.0	0.0	0.0	0.0	0.0	0.0	100.000000	0.000000	0.000000	0.000000
[(0, 0.01252973), (1, 0.01252973), (2, 0.01253...	-0.2	0.4	62.85	...	0.0	0.0	0.0	0.0	0.0	0.0	76.470588	17.647059	66.666667	33.333333
[(0, 0.014307692), (1, 0.15719113), (2, 0.0143...	0.1	0.1	45.42	...	0.0	0.0	0.0	0.0	0.0	0.0	100.000000	0.000000	0.000000	0.000000
[(0, 0.033388544), (1, 0.033388544), (2, 0.033...	0.0	0.0	99.23	...	0.0	0.0	0.0	0.0	0.0	0.0	83.333333	16.666667	50.000000	50.000000

FIGURE 4.4: Dataset after feature extraction

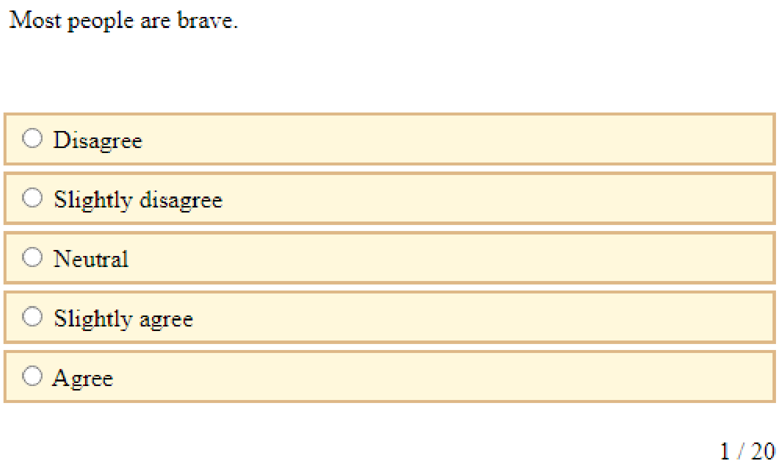
A psychology questionnaire is a form of written self-reporting in which participants are asked to answer a predetermined number of questions. They can be given out in person, via mail, online, on the phone, or concurrently to a number of participants. Numerous psychological dimensions, such as personality, IQ, attitudes, and beliefs, can be evaluated using questionnaires.

Psychology surveys come in a wide variety, each with distinct advantages and disadvantages. The following are some of the most typical questions used in psychology:

Self-report surveys: In these surveys, participants are asked to rate themselves on a range of characteristics, including their personalities, attitudes, and beliefs.

Behavioral checklists: Participants are asked to answer questions about their own behavior, such as how frequently they engage in specific behaviors or how frequently they feel particular emotions. Examining a person's inner ideas, feelings, and motives through the use of ambiguous stimuli like inkblots or pictures is done through projective tests, which ask participants to interpret the information. It is crucial to take into account the precise construct that researchers are interested in testing when selecting a psychological questionnaire. The participants' ages and reading abilities, as well as the time researcher have to deliver the questionnaire, should also be taken into account.

MACH-IV was published to measure this trait in 1970 and it has been a popular instrument ever since and correlates with many things.The test consists of twenty items. Each item is a statement that user must indicate how accurate it would be when applied to the user.[17][18]


Most people are brave.

- Disagree
- Slightly disagree
- Neutral
- Slightly agree
- Agree

1 / 20

FIGURE 4.5: Machiavelli test

Narcissism is basically known as obsession with self love. Narcissists love themselves more than anyone or anything.[19]

Researchers have proposed may questionnaires to detect narcissism in early stage.Many such process is used to help as a tool in psychological inferences.

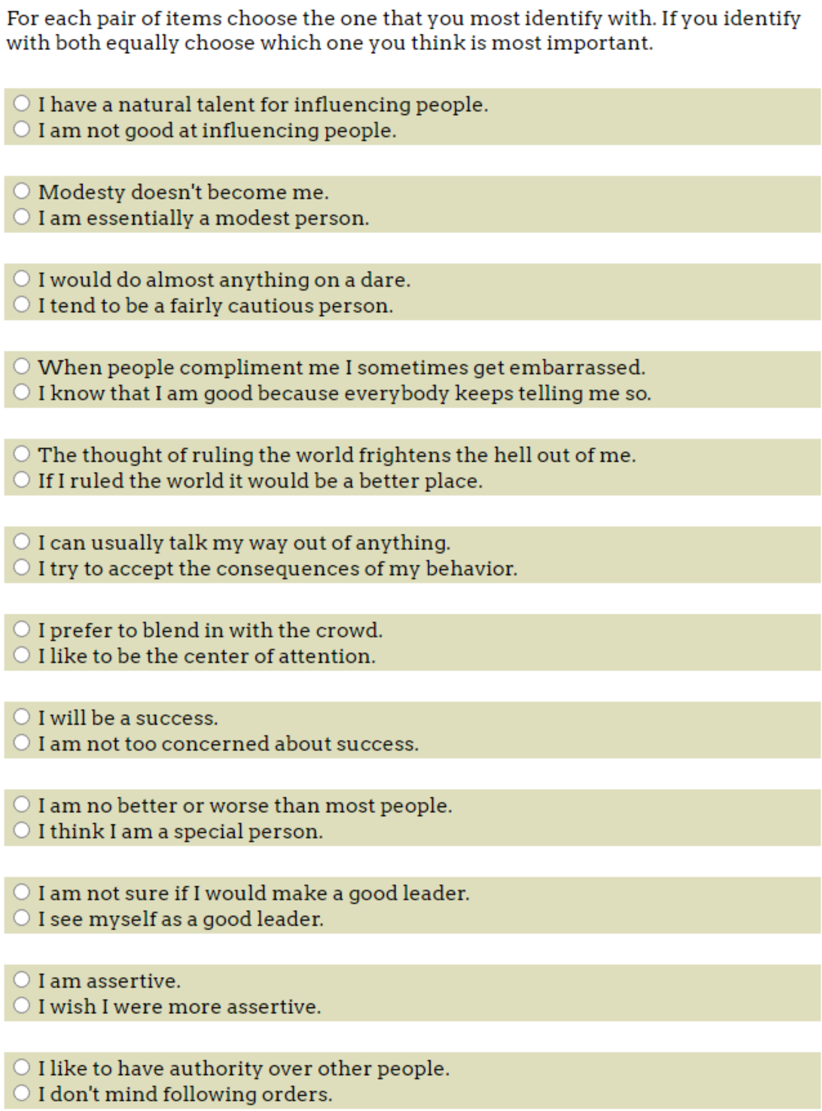
The test was an interactive version of the sociopath test, the Levenson Self-Report Psychopathy Scale.Lack of empathy for others is a trait of psychopathy, a personality condition. In 1995, the LSRP was created for use in psychological research.It measures on two scales: primary psychopathy (psychopathic emotional

affect) and secondary psychopathy (psychopathic lifestyle).The test consists of twenty six statements that could possibly apply to you. You must rate each on how much you agree with it on a scale of (1) strongly disagree (2) disagree (3) neither agree nor disagree (4) agree (5) strongly agree.[20]

We divided the test results by their median after they were taken. This effectively became a binary classification problem since individuals who received results below the median were labeled as having Machiavellian (i.e., if the median for Machiavellian was 60, then people below score 60 was tagged as "lower" for having Machiavellian).

4.1.2 Data pre-processing

As the posts were scraped from facebook and the users are bengali, so, some of the posts were in pure Bangla, where most of them were in English or in banglish form. So, we filtered our dataset.



For each pair of items choose the one that you most identify with. If you identify with both equally choose which one you think is most important.

- ○ I have a natural talent for influencing people.
- ○ I am not good at influencing people.

- ○ Modesty doesn't become me.
- ○ I am essentially a modest person.

- ○ I would do almost anything on a dare.
- ○ I tend to be a fairly cautious person.

- ○ When people compliment me I sometimes get embarrassed.
- ○ I know that I am good because everybody keeps telling me so.

- ○ The thought of ruling the world frightens the hell out of me.
- ○ If I ruled the world it would be a better place.

- ○ I can usually talk my way out of anything.
- ○ I try to accept the consequences of my behavior.

- ○ I prefer to blend in with the crowd.
- ○ I like to be the center of attention.

- ○ I will be a success.
- ○ I am not too concerned about success.

- ○ I am no better or worse than most people.
- ○ I think I am a special person.

- ○ I am not sure if I would make a good leader.
- ○ I see myself as a good leader.

- ○ I am assertive.
- ○ I wish I were more assertive.

- ○ I like to have authority over other people.
- ○ I don't mind following orders.

FIGURE 4.6: Narcissism test

4.1.2.1 Preprocessing for code switching

We analysed the ratio of english and non- english words in a sentence and considered that as a metric of language transformation. We also kept track of code switching points from bangla to english and vice versa.

4.1.2.2 Preprocessing for Feature extraction

For Banglish posts, we will use a converter to convert these posts into pure Bangla, and then we will use a translator to translate into English sentiment. After that, we want to remove special symbols such as “#”, “” and also alpha-numeric symbols from the posts.

	Disagree		Neutral		Agree
Success is based on survival of the fittest; I am not concerned about the losers.					
I find myself in the same kinds of trouble, time after time.					
For me, what's right is whatever I can get away with.					
I am often bored.					
In today's world, I feel justified in doing anything I can get away with to succeed.					
I find that I am able to pursue one goal for a long time.					
My main purpose in life is getting as many goodies as I can.					
I don't plan anything very far in advance.					
Making a lot of money is my most important goal.					
I quickly lose interest in tasks I start.					
I let others worry about higher values; my main concern is with the bottom line.					
Most of my problems are due to the fact that other people just don't understand me.					
People who are stupid enough to get ripped off usually deserve it.					
Before I do anything, I carefully consider the possible consequences.					
Looking out for myself is my top priority.					
I have been in a lot of shouting matches with other people.					

FIGURE 4.7: Psychopathy test

And finally, we will be dealing with emojis. We will convert emojis into plain English texts.

figure 4.8 depicts our data preprocessing task.

4.2 Feature Extraction

As mentioned before, we used hand crafted features and these are some of the features we have used for our research.

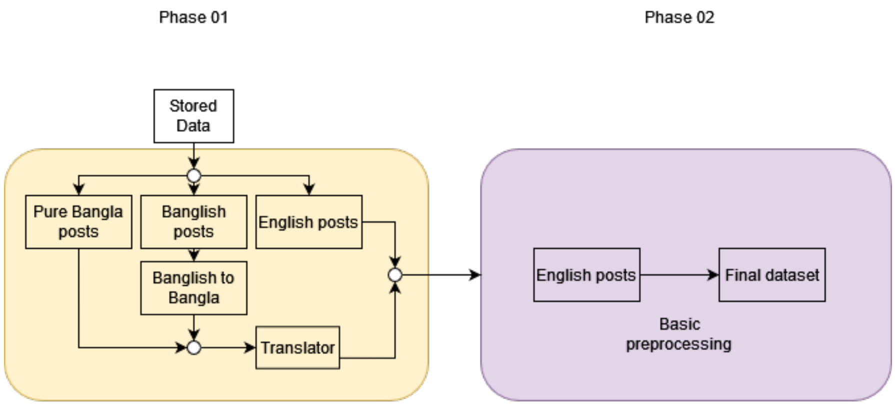


FIGURE 4.8: Data pre-processing

4.2.1 Polarity of text

Polarity refers to the emotional tone of a text, and polarity analysis is a common feature extraction technique used in sentiment analysis. In sentiment analysis, the goal is to determine the sentiment of a piece of text as positive, negative, or neutral. This analysis is typically performed using a machine learning algorithm that has been trained on a labeled dataset of texts with known sentiment. The polarity of a text can be determined by analyzing the sentiment of the words used in the text and how they are used in context.[21]

Polarity analysis is a technique used to determine the overall sentiment of a piece of text. Sentiment can be positive, negative, or neutral. Polarity analysis is often used in sentiment analysis, which is the process of identifying and extracting subjective information from text.[22]

There are a number of ways to perform polarity analysis. One common approach is to use a lexicon, which is a list of words that have been assigned a sentiment score. For example, the word "love" might have a positive sentiment score, while the word "hate" might have a negative sentiment score.[23]

Another approach to polarity analysis is to use a machine learning algorithm. Machine learning algorithms can be trained to identify the sentiment of text based on a labeled data set of text with known sentiment.[23]

Once the polarity of a piece of text has been determined, it can be used to gain insights into the opinions and attitudes of the author or speaker. Polarity analysis can be used in a variety of applications, such as customer feedback analysis, social media monitoring, and political polling.[24]

Here are some examples of how polarity analysis is used:

A company can use polarity analysis to track customer feedback on their products or services. This information can be used to identify areas where customers are happy or unhappy, and to make improvements to the products or services.[25] A political campaign can use polarity analysis to track social media sentiment about their candidate. This information can be used to identify which issues are important to voters, and to tailor the campaign message accordingly. A news organization can use polarity analysis to track sentiment about current events. This information can be used to identify which stories are generating the most interest, and to provide more coverage of those stories. Polarity analysis is a powerful tool that can be used to gain insights into the opinions and attitudes of people. It can be used in a variety of applications, and it is becoming increasingly important as more and more information is being shared online.[26]

4.2.2 Subjectivity

Subjectivity refers to the degree to which a piece of text expresses an opinion or a personal perspective. Subjectivity analysis is used in sentiment analysis to distinguish between objective and subjective text. In subjectivity analysis, a text is classified as subjective if it contains subjective language such as opinions, feelings, or beliefs. This analysis can be performed using machine learning algorithms that are trained on labeled datasets of subjective and objective text.

Subjectivity analysis is the process of determining whether a piece of text expresses an opinion or a personal perspective. This is in contrast to objective text, which simply states facts.

There are a number of ways to perform subjectivity analysis. One common approach is to use a lexicon, which is a list of words that have been classified as subjective or objective. For example, the word ”love” is typically considered to be a subjective word, while the word ”fact” is typically considered to be an objective word.

Another approach to subjectivity analysis is to use a machine learning algorithm. Machine learning algorithms can be trained to identify subjective text based on a labeled dataset of text with known subjectivity.

Once the subjectivity of a piece of text has been determined, it can be used to gain insights into the author’s or speaker’s intentions. Subjectivity analysis can be used in a variety of applications, such as opinion mining, sentiment analysis, and natural language processing.

Here are some examples of how subjectivity analysis can be used:

An opinion mining system can use subjectivity analysis to identify the opinions expressed in a piece of text. This information can then be used to summarize the opinions of a group of people or to identify the most popular opinions. A sentiment analysis system can use subjectivity analysis to identify the sentiment expressed in a piece of text. This information can then be used to determine whether the author or speaker is expressing positive, negative, or neutral sentiment. A natural language processing system can use subjectivity analysis to identify the parts of a sentence that are subjective. This information can then be used to improve the understanding of the sentence by the system. Subjectivity analysis is a powerful tool that can be used to gain insights into the opinions and intentions of people. It can be used in a variety of applications, and it is becoming increasingly important as more and more information is being shared online.

Here are some examples of subjective and objective text:

Subjective text: "I love this movie!" Objective text: "The movie was released in 2019."

4.2.3 Lexical density

Lexical density refers to the proportion of content words (i.e., words that carry meaning) to function words (i.e., words that serve grammatical purposes) in a piece of text. It is a measure of the complexity and richness of the language used in a text. A higher lexical density indicates that the language used in the text is more complex and sophisticated. This feature can be used in text classification tasks to differentiate between texts of varying levels of complexity.

Lexical density is a measure of the complexity and richness of the language used in a text. It is calculated by dividing the number of content words by the total number of words in the text. Content words are words that carry meaning, such as nouns, verbs, adjectives, and adverbs. Function words are words that serve grammatical purposes, such as articles, conjunctions, and prepositions.

A higher lexical density indicates that the language used in the text is more complex and sophisticated. This is because content words are more informative than function words. For example, the sentence "The cat sat on the mat" has a low lexical density because it contains mostly function words. The sentence "The furry feline perched on the woven rug" has a higher lexical density because it contains more content words.

Lexical density can be used in text classification tasks to differentiate between texts of varying levels of complexity. For example, a text classification system might be trained to distinguish between news articles and blog posts. News articles typically have a higher lexical density than blog posts because they are written to inform the reader, while blog posts are often written to express the author’s opinion.

Lexical density can also be used to assess the readability of a text. A text with a high lexical density is likely to be more difficult to read than a text with a low lexical density. This is because content words are often more complex than function words. For example, the sentence ”The cat sat on the mat” is easier to read than the sentence ”The furry feline perched on the woven rug.”

Lexical density is a useful tool for understanding the complexity and richness of language. It can be used in a variety of tasks, such as text classification, readability assessment, and natural language processing.

4.2.4 Word tokens

Word tokens are individual words in a piece of text that are separated by whitespace or punctuation. They are often used as a basic feature in natural language processing tasks. Word tokens can be analyzed to extract features such as word frequency distributions, stopword counts, and part-of-speech tags.

Word tokens are individual words in a piece of text that are separated by whitespace or punctuation. They are often used as a basic feature in natural language processing tasks. Word tokens can be analyzed to extract features such as word frequency distributions, stopword counts, and part-of-speech tags.

Word frequency distributions can be used to identify the most common words in a text. This information can be used to understand the topic of the text or to identify the style of the author. For example, a text about computer science is likely to have a high frequency of words like ”algorithm” and ”programming.” A text about poetry is likely to have a high frequency of words like ”imagery” and ”metaphor.”

Stopword counts can be used to identify words that are not informative. Stopwords are words that are commonly used in everyday language, such as ”the,” ”of,” and ”and.” Stopwords can be removed from a text to improve the performance of natural language processing tasks. For example, a text classification system that is trained on a dataset of news articles is likely to perform better if stopwords are removed from the articles.

Part-of-speech tags can be used to identify the grammatical function of a word. Part-of-speech tags are assigned to each word in a text. For example, the word ”cat” might be tagged as a noun, while the word ”sat” might be tagged as a verb. Part-of-speech tags can be used to understand the structure of a sentence and to identify the relationships between words. For example, a natural language processing system that is trained to parse sentences is likely to perform better if part-of-speech tags are used.

Word tokens are a simple but powerful tool for natural language processing. They can be used to extract a variety of features that can be used to improve the performance of natural language processing tasks.

4.2.5 flesch reading ease

Flesch reading ease is a readability formula developed by Rudolph Flesch in 1948[27]. It is a measure of how easy or difficult a text is to read. The Flesch reading ease score is calculated using the following formula [28]:
Flesch Reading Ease =

$$206.835 - (1.015ASL) - (84.6ASW) \tag{4.1}$$

where
ASL = average sentence length in words
ASW = average number of syllables per word

How simple something is to read is indicated by the Flesch Reading Ease (FRES) score. It is counted by counting the words, syllables, and phrases that are present in the text. The average number of words per phrase and the average number of syllables per word are then calculated. Shorter phrases and words are supposed to be simpler to read. The text is easier to grasp the higher the score. [29] A Flesch reading ease score of 90 or above is considered to be very easy to read, while a score of 30 or below is considered to be very difficult to read. A score of 60 is considered to be easy to read.[30]

The Flesch reading ease score is a useful tool for writers and editors to assess the readability of their writing.[31] It can also be used by teachers and students to assess the readability of textbooks and other materials.

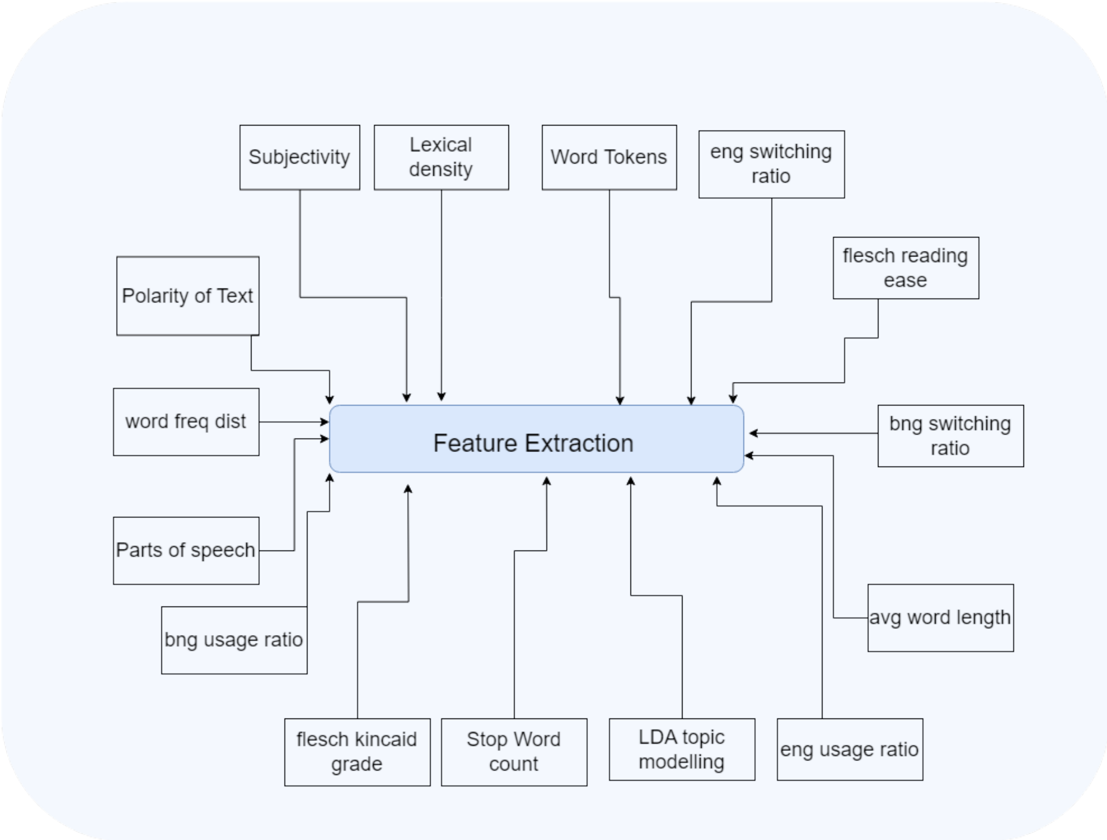


FIGURE 4.9: Extracted Features

4.2.6 flesch kincaid grade

Flesch-Kincaid Grade Level demonstrates the level of education needed to comprehend a text. Text meant for wide public consumption should aim for a grade level of about 8, or 13 to 14.

Flesch kincaid grade =

$$0.39(\frac{totalword}{totalsentence}) + 11.8(\frac{totalsyllables}{totalwords}) - 15.59 \tag{4.2}$$

The results of the two tests correlate almost inversely: a text with a comparatively high score on the Reading Ease test should have a lower score on the Grade-Level test. Rudolf Flesch devised the Reading Ease evaluation; somewhat later, he and J. Peter Kincaid developed the Grade Level evaluation for the United States Navy.

4.2.7 avg word length

The average length per sentence in each post is counted here.

4.2.8 word freq dist

Word frequency distribution is a histogram of the frequency of occurrence of each word in a piece of text. It is often used to identify the most common words in a text and to extract features such as keywords or topics. Word frequency distributions can be analyzed using statistical methods such as Zipf's law to identify patterns in the distribution of word frequencies.

4.2.9 Stop Word count

Stopwords are common words that are often excluded from analysis because they carry little meaning. Stopword count is a feature that counts the number of stopwords in a piece of text. Stopword counts can be used to identify the most common stopwords in a text and to exclude them from analysis.

4.2.10 LDA topic modelling

Latent Dirichlet Allocation (LDA) is a probabilistic model used to identify topics in a collection of texts. LDA topic modeling is a feature extraction technique that identifies the topics present in a piece of text.

4.2.11 Parts of speech

The term "part-of-speech" refers to a property that indicates the grammatical class of each word in a text, such as "noun," "verb," "adjective," etc. Features like named entities or syntactic structures are frequently extracted using this method. Machine learning algorithms that have been trained on labeled datasets of texts using well-known part-of-speech tags can be used to tag parts of speech in texts.

4.2.12 eng usage ratio

The ratio of english words in each post. As our dataset contains both bangla and english words in a single text, so calculated the ratio of bangla and english usage. To do so, we used one dictionary of english words and tagged english words as “eng” words. and remaining words as “bng” words. For a single sentence the ratio was calculated as :

$$eng\_usage\_ratio = \frac{eng\_word\_count}{Total\_words\_count}$$

4.2.13 bng usage ratio

The ratio of bangla words in each post. As our dataset contains both bangla and english words in a single text, so calculated the ratio of bangla and english usage. To do so, we used one dictionary of english words and tagged english words as “eng” words. and remaining words as “bng” words. For a single sentence the ratio was calculated as :

$$bng\_usage\_ratio = \frac{bng\_word\_count}{Total\_words\_count}$$

4.2.14 eng switching ratio

Number of times participant switched from english to bangla language in single text. For instance, “I am going to school” here 0 switching occurred between bangla and english. A flag and count were used to keep track of the position and count of switching between languages. The switching ratio was calculated as following:

$$eng\_switching\_ratio = \frac{number\_of\_switches\_from\_eng\_to\_bangla}{total\_number\_of\_switches}$$

4.2.15 bng switching ratio

Number of times participant switched from bangla to english language in single text. For instance, “I am going to school” here 0 switching occurred between bangla and english. A flag and count were used to keep track of the position and count of switching between languages. The switching ratio was calculated as following:

$$bng\_switching\_ratio = \frac{number\_of\_switches\_from\_bng\_to\_english}{total\_number\_of\_switches}$$

Chapter 5

Experiment

In this chapter, we discuss the experiments we have conducted so far.

5.1 Prediction of Personality Traits

In this section, we represent experiments and their corresponding results for three types of dark personality traits.

5.1.1 Narcissist Prediction

We have used 3 simple machine-learning models such as 1)Naive Bayes 2)SVM 3)RandomForest for predicting narcissistic trait. We also ensembled these models. We have run these models using handcrafted features extracted from texts. We have used 50+ features extracted from text.
Some of the features are: Polarity of text, subjectivity, lexical_density, word_tokens, word_count, avg_word_length, word_freq_dist, stopword_count, part_of_speech, Topic segmentation and so on.
We wanted to try different features to see in which cases the models perform well, so we used "all the features", "features whose zero-sum¿6000" and "features whose zero sum¿4000". we have dropped features whose zero sum¿6000 means, if any feature has 0 values in more than 6000 samples, we dropped that feature column. For example, consider one feature "stopword_count", if it has 0 values in more than 6000 samples, then we dropped that feature. We did this to see if there is any changes or improvements in results.
It mostly occurred in the parts of speech detection part, because the Library that we've used detects 40+ sub-categories of POS that are not significant in usual texts, rather

we focused on the POS which are prominent, by calculating their number of presence in the samples. That's why we used a certain threshold and adjusted it to find necessary parts of speech tags removing rarely occurring ones. We have done these experiments using the code-switching feature, without code-switching features. Code-switching means switching Between different languages in a text. As our dataset contains both Bangla, and English languages, so we calculated the switching ratio between Bangla and English languages for each post. And we considered this as a feature.

5.1.1.1 (a) Using Hand Crafted Features

This table 5.1 represents the results of the experiments using **Naive Bayes** classifier.

Feature Used	**Scaled Dataset**	**Used code-mixing**	**Accuracy**
All	No	No	61.00%
	Yes	No	60.31%
	No	Yes	61.04%
	Yes	Yes	60.43%
Dropped features whose zero sum¿6000	No	No	51.24%
	Yes	No	51.35%
	No	Yes	62.16%
	Yes	Yes	62.45%
Dropped features whose zero sum¿4000	No	No	61.41%
	Yes	No	61.82%
	No	Yes	64.29%
	Yes	Yes	**64.98%**

Table 5.1: Accuracy of Narcissist using naive bayes

This table 5.2 represents the results of the experiments using **SVM** classifier.

Feature Used	**Scaled Dataset**	**Used code-mixing**	**Accuracy**
All	No	No	65.91%
	Yes	No	65.18%
	No	Yes	67.48%
	Yes	Yes	65%
Dropped features whose zero sum¿6000	No	No	66.76%
	Yes	No	65.55%
	No	Yes	67.79%
	Yes	Yes	64.95%
Dropped features whose zero sum¿4000	No	No	67.41%
	Yes	No	65.38%
	No	Yes	**68.54%**
	Yes	Yes	65.44%

Table 5.2: Accuracy of Narcissist using SVM

This table 5.3 represents the results of the experiments using **Random Forest** classifier.

Feature Used	**Scaled Dataset**	**Used code-mixing**	**Accuracy**
All	No	No	75.42%
	Yes	No	75.42%
	No	Yes	75.81%
	Yes	Yes	75.65%
Dropped features whose zero sum¿6000	No	No	79.14%
	Yes	No	78.95%
	No	Yes	88.46%
	Yes	Yes	88.27%
Dropped features whose zero sum¿4000	No	No	75.42%
	Yes	No	80.29%
	No	Yes	90.72%
	Yes	Yes	**90.83%**

Table 5.3: Accuracy of Narcissist using Random Forest

Ensembling of Models: We have ensembled Gaussian Naive Bayes, RandomForest-Classifier, DecisionTreeClassifier, XGBClassifier models. We have used hard voting for ensembling models. The accuracy we have got is: **96.36%** using code-switching as a feature. And without code-mixing we have got **95.21%** accuracy on test set.

5.1.1.2 (b) Using Deep Learning model

We have used **"BERT multilingual model"** for narcissism classification. We used learning_rate $= 2 * e^{-5}$, Number of epochs = 10.
Test set accuracy = **60.34%**

The following table 5.4 represents the classification report of narcissism prediction using BERT multi-lingual model.

	Precision	**Recall**	**F1-score**	**support**
Not Narcissist	0.58	0.84	0.69	447
Narcissist	0.67	0.36	0.47	417
Accuracy			0.61	864
macro avg	0.63	0.60	0.58	864
weighted avg	0.63	0.61	0.58	864

Table 5.4: Classification report

We have also used **"XLM-RoBERTa for sequence classifier** model with number of epoch = 20, Learning rate $= 2^{-5}$. We got test accuracy **49.96%**.

5.1.2 Machiavelli Prediction

We have used 3 simple machine-learning models such as 1)Naive Bayes 2)SVM 3)RandomForest for predicting Machiavelli trait. We also ensembled these models. We have run these models using handcrafted features extracted from texts. We have used 50+ features extracted from text.
Some of the features are: Polarity of text, subjectivity, lexical_density, word_tokens, word_count, avg_word_length, word_freq_dist, stopword_count, part_of_speech, Topic segmentation and so on.
We wanted to try different features to see in which cases the models perform well, so we used "all the features", "features whose zero-sum¿6000" and "features whose zero sum¿4000". we have dropped features whose zero sum¿6000 means, if any feature has 0 values in more than 6000 samples, we dropped that feature column. For example, consider one feature "stopword_count", if it has 0 values in more than 6000 samples, then we dropped that feature. We did this to see if there is any changes or improvements in results.
It mostly occurred in the parts of speech detection part, because the Library that we've used detects 40+ sub-categories of POS that are not significant in usual texts, rather we focused on the POS which are prominent, by calculating their number of presence in the samples. That's why we used a certain threshold and adjusted it to find necessary parts of speech tags removing rarely occurring ones.
We have done these experiments using the code-mixing feature, without code-mixing features. Code-mixing means switching Between different languages in a single sentence. As our dataset contains both Bangla, and English languages, so we calculated the switching ratio between Bangla and English languages for each post. And we considered this as a feature.

5.1.2.1 (a) Using Hand Crafted Features

This table 5.5 represents the results of the experiments using **Naive Bayes** classifier.

Feature Used	**Scaled Dataset**	**Used code-mixing**	**Accuracy**
All	No	No	49.88%
	Yes	No	50.12%
	No	Yes	49.96%
	Yes	Yes	50.06%
Dropped features whose zero sum¿6000	No	No	49.74%
	Yes	No	49.74%
	No	Yes	49.68%
	Yes	Yes	49.63%
Dropped features whose zero sum¿4000	No	No	49.86%
	Yes	No	49.63%
	No	Yes	50.14%
	Yes	Yes	**50.32%**

Table 5.5: Accuracy of Machiavelli using naive bayes

This table 5.6 represents the results of the experiments using **SVM** classifier.

Feature Used	**Scaled Dataset**	**Used code-mixing**	**Accuracy**
All	No	No	59.01%
	Yes	No	58.13%
	No	Yes	59.28%
	Yes	Yes	57.90%
Dropped features whose zero sum¿6000	No	No	58.88%
	Yes	No	58.54%
	No	Yes	**59.92%**
	Yes	Yes	58.05%
Dropped features whose zero sum¿4000	No	No	58.83%
	Yes	No	58.42%
	No	Yes	58.83%
	Yes	Yes	58.37%

Table 5.6: Accuracy of Machiavell using SVM

This table 5.7 represents the results of the experiments using **Random Forest** classifier.

Ensembling of Models: We have ensembled Gaussian Naive Bayes, RandomForestClassifier, DecisionTreeClassifier, XGBClassifier models. We have used hard voting for ensembling models. The accuracy we have got is: **98.27%** using code-mixing as a feature. And without code-mixing we have got **97.85%** accuracy on test set.

5.1.2.2 Using Deep Learning model

We have used **"BERT multilingual model"** for narcissism classification. We used learning_rate $= 2 * e^{-5}$, Number of epochs = 10.

Feature Used	**Scaled Dataset**	**Used code-mixing**	**Accuracy**
All	No	No	72.47%
	Yes	No	72.43%
	No	Yes	75.08%
	Yes	Yes	74.31%
Dropped features whose zero sum¿6000	No	No	75.69%
	Yes	No	76.30%
	No	Yes	83.17%
	Yes	Yes	83.47%
Dropped features whose zero sum¿4000	No	No	76.04%
	Yes	No	76.15%
	No	Yes	**87.79%**
	Yes	Yes	87.54%

Table 5.7: Accuracy of Machiavelli using Random Forest

Test set accuracy = **63.18%**

The following table 5.8 represents the classification report of machiavelli prediction.

	Precision	**Recall**	**F1-score**	**support**
Not Machiavelli	0.59	1.00	0.75	514
Machiavelli	0.00	0.00	0.00	350
Accuracy			0.59	864
macro avg	0.30	0.50	0.37	864
weighted avg	0.35	0.59	0.44	864

Table 5.8: Classification report

We also used **"XLM-RoBERTa for sequence classifier** model with number of epoch = 20, Learning rate= 2^{-5}. We got test accuracy = 62.48%.

5.1.3 Psychopathy Prediction

We have used 3 simple machine-learning models such as 1)Naive Bayes 2)SVM 3)RandomForest for predicting Psychopathy trait. We also ensembled these models. We have run these models using handcrafted features extracted from texts. We have used 50+ features extracted from text.

Some of the features are: Polarity of text, subjectivity, lexical_density, word_tokens, word_count, avg_word_length, word_freq_dist, stopword_count, part_of_speech, Topic segmentation and so on.

We wanted to try different features to see in which cases the models perform well, so we used "all the features", "features whose zero-sum¿6000" and "features whose zero sum¿4000". we have dropped features whose zero sum¿6000 means, if any feature has 0

values in more than 6000 samples, we dropped that feature column. For example, consider one feature ”stopword_count”, if it has 0 values in more than 6000 samples, then we dropped that feature. We did this to see if there is any changes or improvements in results.
It mostly occurred in the parts of speech detection part, because the Library that we’ve used detects 40+ sub-categories of POS that are not significant in usual texts, rather we focused on the POS which are prominent, by calculating their number of presence in the samples. That’s why we used a certain threshold and adjusted it to find necessary parts of speech tags removing rarely occurring ones.
We have done these experiments using the code-mixing feature, without code-switching features. Code-mixing means switching Between different languages in a text. As our dataset contains both Bangla, and English languages, so we calculated the switching ratio between Bangla and English languages for each post. And we considered this as a feature.

5.1.3.1 (a) Using Hand Crafted Features

This table 5.9 represents the results of the experiments using **Naive Bayes** classifier.

Feature Used	Scaled Dataset	Used code-mixing	Accuracy
All	No	No	48.96%
	Yes	No	49.00%
	No	Yes	48.73%
	Yes	Yes	49.00%
Dropped features whose zero sum¿6000	No	No	48.25%
	Yes	No	48.76%
	No	Yes	56.13%
	Yes	Yes	56.24%
Dropped features whose zero sum¿4000	No	No	55.84%
	Yes	No	55.95%
	No	Yes	64.29%
	Yes	Yes	**64.46%**

Table 5.9: Accuracy of Psychopathy using naive bayes

This table 5.10 represents the results of the experiments using **SVM** classifier.

This table 5.11 represents the results of the experiments using **Random Forest** classifier.

Ensembling of Models: We have ensembled Gaussian Naive Bayes, RandomForestClassifier, DecisionTreeClassifier, XGBClassifier models. We have used hard voting for

Feature Used	**Scaled Dataset**	**Used code-mixing**	**Accuracy**
All	No	No	48.50%
	Yes	No	**71.74%**
	No	Yes	51.42%
	Yes	Yes	70.51%
Dropped features whose zero sum¿6000	No	No	45.54%
	Yes	No	71.25%
	No	Yes	50.14%
	Yes	Yes	70.55%
Dropped features whose zero sum¿4000	No	No	42.64%
	Yes	No	70.89%
	No	Yes	50.49%
	Yes	Yes	70.56%

Table 5.10: Accuracy of Psychopathy using SVM

Feature Used	**Scaled Dataset**	**Used code-mixing**	**Accuracy**
All	No	No	95.13%
	Yes	No	95.16%
	No	Yes	95.74%
	Yes	Yes	95.71%
Dropped features whose zero sum¿6000	No	No	96.86%
	Yes	No	96.63%
	No	Yes	99.27%
	Yes	Yes	**99.28%**
Dropped features whose zero sum¿4000	No	No	95.13%
	Yes	No	97.24%
	No	Yes	99.19%
	Yes	Yes	99.16%

Table 5.11: Accuracy of Psychopathy using Random Forest

ensembling models. The accuracy we have got is: **99.62%** using code-mixing as a feature. And without code-mixing we have got **99.46%** accuracy on test set.

5.1.3.2 Using Deep Learning model

We have used **"BERT multilingual model"** for narcissism classification. We used learning_rate $= 2 * e^{-5}$, Number of epochs = 10.
Test set accuracy = **62.76%**

The following table 5.12 represents the classification report of psychopath prediction.

We also used **"XLM-RoBERTa for sequence classifier** model with number of epoch = 20, Learning rate= 2^{-5}. We got test accuracy = %.

	Precision	**Recall**	**F1-score**	**support**
Not Psychopath	0.62	0.85	0.72	479
Psychopath	0.66	0.36	0.47	385
Accuracy			0.63	864
macro avg	0.64	0.61	0.59	864
weighted avg	0.64	0.63	0.61	864

TABLE 5.12: Classification report

5.2 Simple RNN to predict the personality from the labeled dataset

```
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding_7 (Embedding)     (None, 4587, 200)         1880000

 simple_rnn_7 (SimpleRNN)    (None, 32)                7456

 dense_14 (Dense)            (None, 10)                330

 dense_15 (Dense)            (None, 1)                 11

=================================================================
Total params: 1,887,797
Trainable params: 1,887,797
Non-trainable params: 0
_________________________________________________________________
```

We used simple rnn to detect dark traid from the dataset. The performance was very poor after some epochs.

Cause of poor performance: When the users gave the SD3 test, they were given scores on their personality. Next we considered the highest score of his among the three traits as his personality. For example, a person has score highest in narcissism among all, then he is considered narcissist. But this is not the case actually. Only some of his posts maybe narcissistic but others are not. We did not correct this information in our dataset. It resulted in a particular bias in case of all our users.

```
Epoch 1/5
68/68 [==============================] - 219s 3s/step - loss: 0.6198 - accuracy: 0.6782 - val_loss: 0.6054 - val_accuracy: 0.7035
Epoch 2/5
68/68 [==============================] - 221s 3s/step - loss: 0.4291 - accuracy: 0.8165 - val_loss: 0.6246 - val_accuracy: 0.6667
Epoch 3/5
68/68 [==============================] - 229s 3s/step - loss: 0.1950 - accuracy: 0.9451 - val_loss: 0.7121 - val_accuracy: 0.6427
Epoch 4/5
68/68 [==============================] - 221s 3s/step - loss: 0.0642 - accuracy: 0.9876 - val_loss: 0.8294 - val_accuracy: 0.6096
Epoch 5/5
68/68 [==============================] - 222s 3s/step - loss: 0.0272 - accuracy: 0.9949 - val_loss: 0.9216 - val_accuracy: 0.5985
```

5.3 Sentiment Analysis

In this section, we have discussed how we have analysed the semtiments of every individual based on his social media posts.

5.3.1 Bert model to classify the sentiment of a sentence trained on IMDB dataset

For this experiment, we used BERT model to classify the sentiment of each post. And then we analyzed the fluctuation of the sentiment for a specific person for a period of time. Here time range of the posts of each person varies. Some posts are consecutive, some with big time gaps. Moreover, some users may have 100 posts in a year whereas others may have 100 posts in two to three years. So, the sentiment fluctuation of each person also differs from each other.

posts	Mach	LSRP12	LSRP2	NRSM	languages	sentiment
my heart my soul..T-T	68.0	1.3	2.5	3.0	en	Positive
Free speech is a joke in this country	68.0	1.3	2.5	3.0	en	Negative
Make 500 sequels I don't care.Make another Jus...	68.0	1.3	2.5	3.0	en	Negative
Everybody wants a piece of the politically cor...	68.0	1.3	2.5	3.0	en	Negative
And that's how a movement dies.	68.0	1.3	2.5	3.0	en	Negative
...	...	...	...	...	...	...
This broke my heart... May the departed soul r...	53.0	1.9	3.1	19.0	en	Positive
:(First Massive crash with Bangladeshi passen...	53.0	1.9	3.1	19.0	en	Positive
TWO SUPERHUMANS IN ONE FRAME <3	53.0	1.9	3.1	19.0	en	Negative
Man!!!\nThe chase- UNBELIEVABLE!! Mushi and Ri...	53.0	1.9	3.1	19.0	en	Negative
A small piece of heaven on earth.	53.0	1.9	3.1	19.0	en	Positive

FIGURE 5.1: Predicted sentiment of the posts

First, we finetuned the BERT model and then trained the model on IMDB dataset. This dataset is a labeled dataset that has two labels: 1)positive sentiment and 2)negative sentiment. After training the model, we used our dataset as test data and then found the sentiment for each of the posts, see figure 5.1 for reference.

Finally, we plot the sentiment of a specific person over the posts. The graph shows the rate of fluctuation of a person fig 5.2.

The objective of this experiment: We wanted to know the rate at which a person's sentiment changes. If the sentiment of a person changes frequently then we could consider that is not normal behavior. For example, narcissists do not have much fluctuation in their sentiment visually but psychopaths do. Psychopaths have a wide range mood swing which shows very often with their disappointment in any situation. And it may have some relationship with dark triad traits.

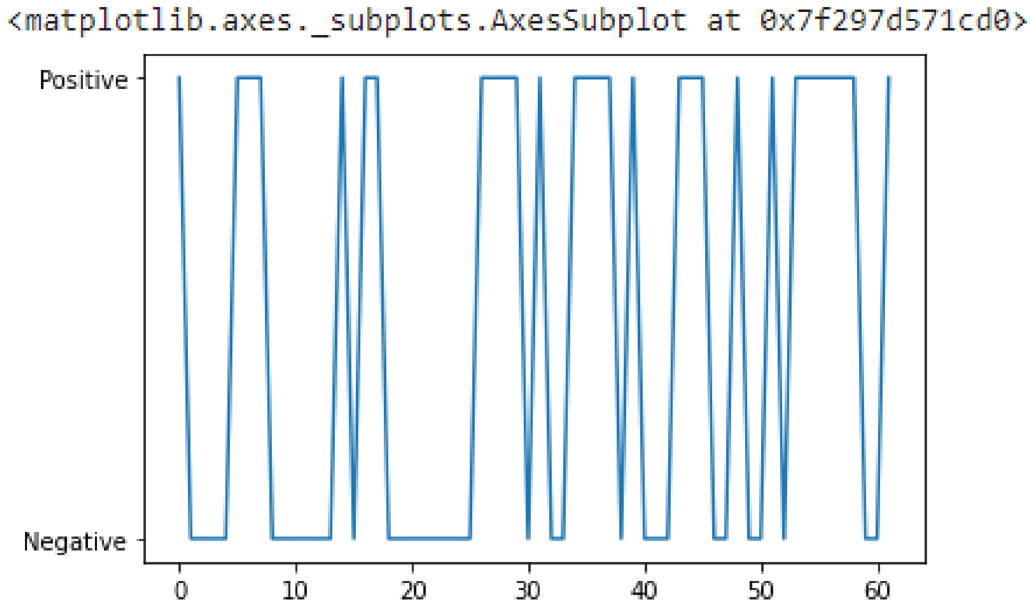


Figure 5.2: Fluctuation of sentiment of a person over posts

Problem: It has many problems. First of all, the model we used was trained on IMDB dataset which is a movie ratings dataset and also it only contains positive and negative sentiment, which does not match to our labels (psychopathy, machiavellianism, narcissism). Again, fluctuation of sentiment may not imply any relationship with dark triad. Also the number of posts is very low which may not reflect the real value of dark triad.

5.3.2 Bert model to classify the sentiment of a sentence trained on DistilBERT

We did the above-mentioned sentiment classification graph using another pre trained model. This model is a fine-tuned checkpoint of DistilBERT-base-uncased, fine-tuned on SST-2. This model reaches an accuracy of 91.3 on the dev set (for comparison, the Bert bert-base-uncased version reaches an accuracy of 92.7). This was trained on tweeter data so relevant to our experiment, unlike the previously used IMDB dataset. The level was ranging between 0 to 1 reflecting the intensity of the emotion.

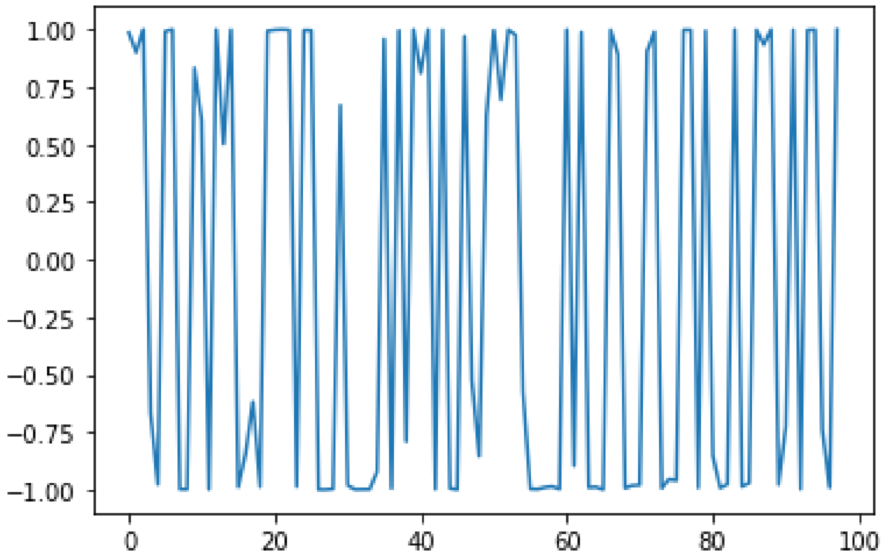


FIGURE 5.3: Fluctuation of sentiment of a person over posts

5.3.3 DistilBert model to classify the emotion of a sentence trained on emotion dataset

For this we have used DistilBert model which was trained on "emotion" dataset. This dataset had 6 emotions such as: anger, sadness, joy, fear, surprise and love. We have used this model to predict the emotion of the posts. Inclusion of each emotion leads to different personalities of people. The emotions can be mututally inclusive in personality which makes it harder to distinguish between particular personality symptoms. The following figure 5.4 shows the sentiment and score for that sentiment added to the original dataset.

nnamed: 0	index	Unnamed: 0.1	posts	Mach	LSRP12	LSRP2	NRSM	languages	Length	sentiment	person	Person	sentiment score
0	0	0	my heart my soul..T-T	68.0	1.3	2.5	3.0	en	21	anger	2	2	0.839829
1	2	2	Free speech is a joke in this country	68.0	1.3	2.5	3.0	en	37	joy	2	2	0.628497
2	5	5	Make 500 sequels I don't care.Make another Jus...	68.0	1.3	2.5	3.0	en	110	sadness	2	2	0.973745
3	6	6	Everybody wants a piece of the politically cor...	68.0	1.3	2.5	3.0	en	55	joy	2	2	0.992939
4	10	10	And that's how a movement dies.	68.0	1.3	2.5	3.0	en	31	sadness	2	2	0.975138

FIGURE 5.4: Fluctuation of emotion of a person over posts

After that, we have plotted the emotions over posts for each person. It shows the fluctuation of sentiments in each person based on his regular posts from social media. For instance, the graph for person 2 is shown in figure 5.5

```
sns.lineplot(data=frequency_data[2])
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f374e79f280>
```

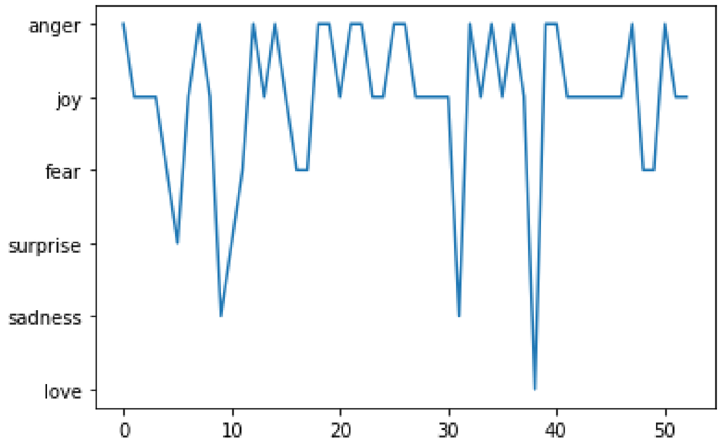


FIGURE 5.5: Fluctuation of emotion of a person over posts

The following figure 5.6 shows the count of posts for each of the emotions.

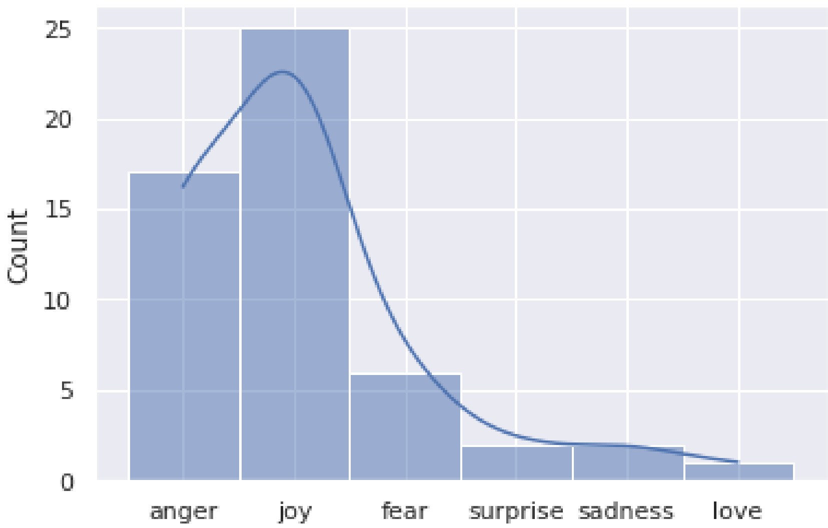


FIGURE 5.6: Fluctuation of the emotions of a person over posts

Chapter 6

Analysis

In this section, we will discuss and compare the results of different models performed on our dataset with other existing datasets. To compare our result with existing results we will consider the features they have used. We have already shown that machine learning models have outperformed the deep learning models.

6.1 Narcissist

To predict narcissist characteristics of the posts we have used Naive Bayes, SVM, Random Forest and emsembled models. This paper [2] predicts narcissist triat of posts collected from twitter. They have used Naive Bayes and SVM models for this purpose. They have used "Bag of Words" model to quickly identify words related to a specific domain. They have got 0.88 as the highest F1 score and accuracy was 0.88 . For our case, we have extracted 50+ features and got 0.96 as the highest F1 score by ensembled 3 models. And highest accuracy was 96.36%.

Paper [32] used a dataset that was originally developed by this [7] paper using twitter posts. They have used 224 features, including analyzed text features (unigrams, word frequencies, word clusters) sentiments, user demographics, platform usage, and profile images features for the prediction of the Dark Triad. They have used Linear Regression and Random Forest models for dark triad prediction. They have got highest accuracy as 69.2%, whereas we have used 50+ features including Polarity of text, subjectivity, lexical_density, word_tokens, word_count, avg_word_length, word_freq_dist, stopword_count, part_of_speech, Topic segmentation and so on. And we have got highest accuracy as 96.36% by ensembling 3 ML models.

6.2 Psychopath

To predict narcissist characteristics of the posts we have used Naive Bayes, SVM, Random Forest and emsembled of three models. Paper [4] used BiLSTM deep learning model to detect psychopathy characteristics from social media posts. They have got test accuracy 85% and F1 score as 0.85. But we using machine learning model we got highest accuracy of 99.61% and F1 score is 1.00, which outperforms the result of deep learning model.

Paper [32] got 75.7% accuracy using Linear Regression model, using the same set of features discussed in section 6.1. For our case we got 99.61% accuracy by ensembling 3 ML models.

6.3 Machiavelli

Paper [32] used same set of features and got 73.8% as the highest accuracy. On the other hand, we got 98.27% accuracy as highest accuracy.

Chapter 7

Futute Works

For the future work, we want to work with emojis of the texts. If emoji_count or emotion of emoji are an important feature for detecting dark triad trait of a person using social media posts.

We want to collaborate with psychologists for finding dominant features of persons with dark triad characteristics. For instance, if word_length feature is a dominant feature for persons with machiavelli traits will be verified by professional psychologists. This can be a novel contribution in the field of dark triad detection.

We want to collect more data in future and we want to remove informal language (Banglish) from the dataset. We want to observe that if doing these increases the accuracy of transformers based language models.

Bibliography

[1] F. große Deters, M. R. Mehl, and M. Eid, "Narcissistic power poster? on the relationship between narcissism and status updating activity on facebook," *Journal of Research in Personality*, vol. 53, pp. 165–174, 2014.

[2] L. Haz, M. Á. Rodríguez-García, and A. Fernández, "Detecting narcissist dark triad psychological traits from twitter." in *ICAART (2)*, 2022, pp. 313–322.

[3] E.-M. Rathner, J. Djamali, Y. Terhorst, B. Schuller, N. Cummins, G. Salamon, C. Hunger-Schoppe, and H. Baumeister, "How did you like 2017? detection of language markers of depression and narcissism in personal narratives," 2018.

[4] J. Asghar, S. Akbar, M. Z. Asghar, B. Ahmad, M. S. Al-Rakhami, and A. Gumaei, "Detection and classification of psychopathic personality trait from social media text using deep learning model," *Computational and Mathematical Methods in Medicine*, vol. 2021, 2021.

[5] F. Alotaibi, D. M. Asghar, and S. Ahmad, "A hybrid cnn-lstm model for psychopathic class detection from tweeter users," *Cognitive Computation*, vol. 13, 05 2021.

[6] M. Hassanein, S. Rady, W. Hussein, and T. Gharib, "Predicting the dark triad for social network users using their personality characteristics," *International Journal of Computers and Their Applications*, vol. 28, pp. 204–211, 12 2021.

[7] D. Preotiuc-Pietro, J. Carpenter, S. Giorgi, and L. Ungar, "Studying the dark triad of personality through twitter behavior," in *Proceedings of the 25th ACM international on conference on information and knowledge management*, 2016, pp. 761–770.

[8] D. L. Paulhus and K. M. Williams, "The dark triad of personality: Narcissism, machiavellianism, and psychopathy," *Journal of research in personality*, vol. 36, no. 6, pp. 556–563, 2002.

[9] J. A. Golbeck, "Predicting personality from social media text," *AIS Transactions on Replication Research*, vol. 2, no. 1, p. 2, 2016.

[10] V. Kharde, P. Sonawane *et al.*, "Sentiment analysis of twitter data: a survey of techniques," *arXiv preprint arXiv:1601.06971*, 2016.

[11] Y. Mehta, N. Majumder, A. Gelbukh, and E. Cambria, "Recent trends in deep learning based personality detection," *Artificial Intelligence Review*, vol. 53, pp. 2313–2339, 2020.

[12] B. Pang and L. Lee, "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts," *arXiv preprint cs/0409058*, 2004.

[13] H. Zheng and C. Wu, "Predicting personality using facebook status based on semi-supervised learning," in *Proceedings of the 2019 11th International Conference on Machine Learning and Computing*, 2019, pp. 59–64.

[14] S. M. Bergman, M. E. Fearrington, S. W. Davenport, and J. Z. Bergman, "Millennials, narcissism, and social networking: What narcissists do on social networking sites and why," *Personality and individual differences*, vol. 50, no. 5, pp. 706–711, 2011.

[15] M. Conway and D. O'Connor, "Social Media, Big Data, and Mental Health: Current Advances and Ethical Implications," *Current Opinion in Psychology*, vol. 9, pp. 77–82, Jun. 2016.

[16] R. Landers, R. Brusso, K. Cavanaugh, and A. Collmus, "A Primer on Theory-Driven Web Scraping: Automatic Extraction of Big Data From the Internet for Use in Psychological Research," *Psychological Methods*, vol. 21, May 2016.

[17] R. Christie, "CHAPTER XVI - SOCIAL CORRELATES OF MACHIAVELLIANISM," in *Studies in Machiavellianism*, R. Christie and F. L. Geis, Eds. Academic Press, Jan. 1970, pp. 314–338. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B978012174450250021X

[18] R. Christie and F. Geis, "CHAPTER XVII - IMPLICATIONS AND SPECULATIONS," in *Studies in Machiavellianism*, R. Christie and F. L. Geis, Eds. Academic Press, Jan. 1970, pp. 339–358. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780121744502500221

[19] R. Raskin and H. Terry, "A principal-components analysis of the Narcissistic Personality Inventory and further evidence of its construct validity," *Journal of Personality and Social Psychology*, vol. 54, no. 5, pp. 890–902, May 1988.

[20] M. R. Levenson, K. A. Kiehl, and C. M. Fitzpatrick, "Assessing psychopathic attributes in a noninstitutionalized population," *Journal of Personality and Social Psychology*, vol. 68, no. 1, pp. 151–158, Jan. 1995.

[21] M. A. Cabanlit and K. J. Espinosa, "Optimizing N-gram based text feature selection in sentiment analysis for commercial products in Twitter through polarity lexicons," in *IISA 2014, The 5th International Conference on Information, Intelligence, Systems and Applications*, Jul. 2014, pp. 94–97.

[22] A. Sharma and U. Ghose, "Sentimental Analysis of Twitter Data with respect to General Elections in India," *Procedia Computer Science*, vol. 173, pp. 325–334, Jan. 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050920315428

[23] B. Brahimi, "Data and Text Mining Techniques for Classifying Arabic Tweet Polarity," vol. 14, no. 1, 2016.

[24] Z. Ren, Q. Shen, X. Diao, and H. Xu, "A sentiment-aware deep learning approach for personality detection from text," *Information Processing & Management*, vol. 58, no. 3, p. 102532, May 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306457321000406

[25] M. M. Mostafa, "An emotional polarity analysis of consumers' airline service tweets," *Social Network Analysis and Mining*, vol. 3, no. 3, pp. 635–649, Sep. 2013. [Online]. Available: https://doi.org/10.1007/s13278-013-0111-2

[26] Y. Niu, X. Zhu, J. Li, and G. Hirst, "Analysis of Polarity Information in Medical Text," *AMIA Annual Symposium Proceedings*, vol. 2005, pp. 570–574, 2005. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1560818/

[27] R. Flesch, "A new readability yardstick," *Journal of Applied Psychology*, vol. 32, pp. 221–233, 1948, place: US Publisher: American Psychological Association.

[28] J. N. Farr, J. J. Jenkins, and D. G. Paterson, "Simplification of Flesch Reading Ease Formula," *Journal of Applied Psychology*, vol. 35, pp. 333–337, 1951, place: US Publisher: American Psychological Association.

[29] D. Eleyan, A. Othman, and A. Eleyan, "Enhancing Software Comments Readability Using Flesch Reading Ease Score," *Information*, vol. 11, no. 9, p. 430, Sep. 2020, number: 9 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/2078-2489/11/9/430

[30] A. R. Sutin, P. T. Costa Jr., M. K. Evans, and A. B. Zonderman, "Personality assessment in a diverse urban sample," *Psychological Assessment*, vol. 25, pp. 1007–1012, 2013, place: US Publisher: American Psychological Association.

[31] "The Influence of Personality Traits on Narrative Writing Skills in Third and Fourth-Grade Students - ProQuest." [Online]. Available: https://www.proquest.com/openview/279715ee92d9d1f8c066f1b378c76a52/1?pq-origsite=gscholar&cbl=18750&diss=y

[32] M. Hassanein, S. Rady, W. Hussein, and T. F. Gharib, "Predicting the dark triad for social network users using their personality characteristics." *International Journal for Computers & Their Applications*, vol. 28, no. 4, 2021.