Arab American University

Faculty of Graduate Studies

Mining X Platform Data for Predictive Personality Modeling

By

Tayseer Ahmad Mohammad Othman

Supervisor

Dr. Sami Sader

This thesis was submitted in partial fulfillment of the requirements for the Master's degree in Data Science and Business Analytics

February/ 2024

© Arab American University – 2024. All rights reserved.

Thesis Approval

Mining X Platform Data for Predictive Personality Modeling.

By

Tayseer Othman

This thesis was defended successfully on 27.2.2024 and approved by:

Committee members	Signature
1. Dr. Sami Sader: Supervisor	
2. Dr. Majdi Owda: Internal Examiner	
3. Dr. Radi Jarrar; External Examiner	

Declaration

I declare that this thesis has been composed solely by myself and that it has not been submitted, in whole or in part, in any previous application for a degree. Except where stated otherwise by reference or acknowledgment, the work presented is entirely my own.

Candidate Name: Tayseer Ahmad Mohammad Othman

Student ID: 201912678

Signature: Tayseer Othman

Date: 17.9.2025

Dedication

To my dearest parents,

For your unwavering love, encouragement, and sacrifices, which have been my guiding light throughout this journey. Your boundless support has made this achievement possible.

To my beloved wife,

For your patience, understanding, and the countless sacrifices you made to support me in this academic endeavor. You are my pillar of strength and the reason behind my success.

To my precious children, Mohammad, Farah, and Ahmad,

Your innocent smiles and boundless joy have been my motivation. I dedicate this thesis to you, hoping to be a role model and source of inspiration as you grow and learn about the beauty of knowledge and determination.

With love and gratitude,

Tayseer Othman

Acknowledgment

I would like to express my heartfelt gratitude to the many individuals whose support and contributions have made the completion of this thesis possible. This journey of exploring the complex world of mining user personality on X Platform has been both challenging and enlightening, and I am deeply appreciative of the following:

First and foremost, I extend my profound thanks to my academic advisor Dr. Sami Sader, whose guidance, wisdom, and unwavering support have been instrumental throughout this research. Your mentorship and dedication to fostering intellectual growth have been invaluable.

I am thankful to the faculty and staff at Arab American University Palestine, whose commitment to academic excellence and research opportunities has provided a conducive environment for scholarly endeavors.

I owe a debt of gratitude to my family for their enduring encouragement, patience, and understanding. Your belief in me and your constant motivation have been my pillars of strength.

I extend my love and appreciation to my friends and colleagues who have provided both academic and emotional support. Your discussions and perspectives have enriched this work.

This thesis is a testament to the collective efforts and support of the aforementioned individuals, and it would not have been possible without each of you.

Thank you.

Abstract

Social media has gained widespread acceptance and is now a prominent feature of people's modern digital landscape. A significant portion of human endeavors, including interactions with others, entertainment consumption, online shopping, and information seeking, are increasingly conducted through digital platforms and devices. Previous studies have revealed that more than 2.4 billion people of all ages use social media, and they spend a significant amount of their time engaged in it. Engaging in actions such as updating profiles, posting status updates, and sharing various content offers users a means to express a substantial amount about themselves. They showcase their personalities by using self-descriptions, sharing their current statuses, uploading pictures, and highlighting their interests. Social media profiles and tweets frequently serve as channels through which users provide glimpses into their traits and characteristics.

This study addresses the predictive capacity of social media profiles regarding personality traits. This research work analyzed the data generated from interactions on social media platforms, intending to uncover connections and associations between online actions and diverse personality traits. Data analysis methodologies such as natural language processing and deep learning methods are utilized to examine elements such as language utilization, content inclinations, and interaction tendencies. Through these techniques, the study seeks to understand potential indicators of personality traits.

In this study, utilizing deep learning algorithms such as Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRUs) models after cleaning a dataset of X Platform posts. We found that these tools can be utilized to specify users' personalities by analyzing the Tweets users' posts. A 90% accuracy of predicting user personality was achieved. Consequently, this

project is evidence of the possibility of utilizing social media data mining to reach more accurate results and predict the changes in the users' personalities.

Keywords: social media, machine learning, personality, prediction, characteristics, and crossplatform, self-descriptions.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

Abbreviation	Expansion
MBTI	Myers-Briggs Type Indicator
API	Application Program Interface
NaN	Not A valid Number
LSTM	Long Short-Term Memory
NLP	Natural Language Processing
HR	Human Resources
LIWC	Linguistic Inquiry and Word Count
RBF	Radial Bases Function
ID3	Iterative Dichotomiser 3
K-NN	k-Nearest Neighbors
SVM	Support Vector Machine
XgBoost	A gradient-boosting algorithm
POS	Part of Speech
WEKA	Waikato Environment for Knowledge Analysis
SMOTE	Synthetic Minority Oversampling TEchnique
IWVTSA	Improved Word Vectors for X Platform Sentiment Analysis
VADER	Valence Aware Dictionary and sEntiment Reasoner
RT	retweet
NLTK	Natural Language Toolkit
SIA	Sentiment Intensity Analyzer
GRU	Gated Recurrent Unit
DL	Deep Learning
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
ROC	Receiver Operating Characteristic

EDA	Exploratory Data Analysis

Chapter One: Introduction

1.1 Overview

Social media platforms strongly affect how people live each day. This established a great opportunity for data scientists to study them over time so they can be utilized for selling goods and things with stronger advertising with the ability to know and classify people according to their political opinions and views. Data scientists now can monitor closely people's actions over social media by classifying, clustering, or categorizing what kind of personalities they have. X platform (previously known as Twitter) is a popular platform around the world. Millions of people use it daily. So, it is a huge place for researchers.

This research uses data science techniques to analyze users' posts over X platform by studying the X platform users' emotional status, collecting the common things between their posts to understand the relationships between their posts, and the ability to categorize their personalities. With the use of advanced technology available easily in the market, social media has become a very important resource in the fields of data science and data analysis. Through Twitter, insights about peoples' expressions, feelings, and emotions can be gained through their posts, which may reveal their psychological situations. The analysis of text in the posts or the use of Deep Learning techniques can potentially uncover the emotions and feelings of users on this platform.

The significance of this research extends to decision-makers, steering committees, commentators, and large companies, as it can aid in managing social networks and understanding and analyzing emotional aspects. Additionally, the study sheds light on psychological issues that some social media users may experience, such as Cyberbullying, Psychopathy, and Narcissism.

In conclusion, this study helps to understand how people act on social media platforms such as Twitter. In addition, the results of this study can help in figuring out emotions, promoting mental health awareness, and managing social media content.

1.2 Social Media Platforms

Many people use social media platforms such as Facebook, X, Instagram, WhatsApp, Messenger (Meta), Snapchat, TikTok, and others. Additionally, people seeking professional networking can use LinkedIn as the world's leading professional networking platform. People use these platforms every day to connect with friends and different kinds of people. They talk about different topics on social media, like religion, what they know, what they like, their political thoughts, and what they do (Boyd & Ellison, 2007). According to Kepios (2022), 4.7 billion people use social media, approaching about 59% of all people on Earth. People usually go to around 7.4 different social media sites each month and spend about 2.5 hours each day using these platforms.

1.3 Problem Statement

The exploration of whether social media profiles can accurately predict personality traits significant importance in the digitized world of today (Watt et al., 2023). As social media platforms have become widely spread over the world and deeply ingrained in people’s daily routines, they serve as a treasure trove of valuable data about individual behavior, preferences, and interactions. According to Stacy Jo Dixon (2023), With over 3 billion people using platforms like Facebook across diverse age groups, it's evident that these digital spaces have a profound impact on how users communicate, express themselves, and engage with the world.

Different activities on social media platforms as X platform for changing and modifying different profile contents offer rich information and guidance to users' personality

identifications and preferences. With the easy use and the ability of content sharing and presenting, users unintentionally spread fingerprints and provide a digital mirror of their personality traits. Online activities and behaviors such as writing posts consisting of words, image uploads and likes with users' interests can strongly reflect complicated aspects of the users' psychological aspects (Rijitha R, 2023).

This research aims to dive deeply into the aspects of the X platform users' digital footprints and marks. This research aims to understand patterns and connections to get X platform users' personalities and online activities by exploiting advanced machine learning and natural language processing technologies. The purpose of this study and research is to uncover the relationship of hidden aspects that can't be apparent to the human mind. The ultimate achievement of this research is to be able to identify activities and behaviors that can be treated as a reliable indicator of the users' personality traits.

The study has wider results and outcomes for many fields. For psychology, it provides new insights into how users interact with subjects and communicate their identities in the digital space. This knowledge might completely change marketing by enabling the material to be customized to the innate preferences of each individual. Insights on how online activities and interactions mirror users' social, cultural, and behavioral behaviors in online actions.

Additionally this study effort offers a view into the complex relationship between data science technologies and human psychology with the ability to change people's perceptions and the interaction between virtual space and real-world environment as the digital era changes and data analysis and analytics tools are improving day by day.

1.4 Research Objectives

This research aims to address the possibility of understanding the personality of users by analyzing their tweets on X Platformusing data mining techniques. This will include data collection, preparation, analysis, and deep learning to predict personality traits.

Based on this goal, the following objectives are set:

1. **Data preparation and pre-processing**: obtain the suitable dataset including tweets that are gathered publicly and are available for scientific research. This goal includes data optimization and preparation.
2. **Utilizing data mining techniques to analyze data:** in this project, Natural Language Processing (NLP) and Deep Learning algorithms are utilized to analyze the gathered and preprocessed dataset in order to identify patterns and extract users' indicative personality traits features.
3. **Validating the accuracy and reliability of the predictive models:** this will be done by utilizing accuracy analysis on the obtained results and comparing the training dataset results with the testing dataset results.
4. **Utilize the resulting model to uncover the relationship between tweets and personality traits** by identifying the linguistic cues and behavioral patterns in tweets, and correlating them with different personality traits. The aim is to gain deeper insights into the relationship between language used and personality expression on Twitter.
5. **Evaluate the practical applicability of research findings on other potential realworld areas**, such as personalized content recommendations, user profiling, and targeted marketing strategies.

6. **Discussing the ethical and privacy concerns** linked to the topic of this study and the constraints of predicting personality traits using publicly accessible data from Twitter. The aim is to ensure ethical standards and safeguard user privacy commitment while sharing personal things online.

Through achieving these goals, the thesis seeks to provide important insights into the realm of predicting personality traits using X Platform data analysis. This involves understanding the viability and precision of these techniques and their potential implications across fields like psychology, social sciences, and digital marketing.

1.5 Questions of the Study

The primary question of this study is about the possibility of utilizing data science approaches and methods such as deep learning in natural language processing NLP, to predict personality traits by analyzing public posts that are published on one of the popular social media platforms such as X platform.

Accordingly, the sub-questions further delve into the main question of the research:

1. Do Twitter's posts identify relevant indicators for users' personalities?

This question focuses on determining if the content shared by users on X Platformholds significant signals that can be employed to forecast an individual's personality traits. It involves analyzing the language, sentiments, and topics in the posts to identify patterns related to personality.

2. Do social media users present a list of indicators?

In this sub-question, the study seeks to provide a comprehensive list of indicators extracted from the X Platform data that are strongly associated with specific personality traits. This list can serve as a reference for future studies and applications.

3. What are the shortcomings of the effectiveness of different classification methods for predicting personality traits?

This question explores the limitations and challenges encountered when using different classification techniques to predict personality traits. Understanding these shortcomings can guide researchers in selecting appropriate methods and improving accuracy.

4. What is the feasibility of predictive personality modeling?

The study aims to assess the practicality and reliability of using social media data, particularly Twitter, to predict personality traits accurately. This includes evaluating the effectiveness of various modeling approaches and their real-world applicability.

5. What are the factors that affect personality?

This sub-question addresses the broader factors that influence an individual's personality, beyond just social media usage. It involves exploring psychological, environmental, and social factors that play a role in shaping a person's personality traits.

The thesis will involve data collection from X Platform profiles and posts, applying data mining and machine learning techniques for analysis, and evaluating the results based on established personality assessments. By addressing these research inquiries, the study aims to add to the developing domain of forecasting personality traits and provide insights into the possible outcomes and uses of analyzing social media to understand human behavior and personality traits.

1.6 Research Propositions

1. To recognize the significant indicators linked to personality traits that are clear, either directly or indirectly, in the X Platform posts of user data.
2. To investigate the practicality and feasibility of using predictive personality modeling to enhance upcoming intelligence systems.

3. Investigate and enhance the effectiveness of different classification methods for modeling personality traits.

1.7 Importance of the Study

Since people's personalities uniquely shape who they are and impact how they behave, feel, and think, there's a big opportunity to add new personality-based features to the ways of designing user interfaces. In areas like online shopping, working together, online learning, and organizing information, personalized systems can benefit from user interfaces that adapt to how each person interacts. This involves adjusting things like how you interact, what motivates you, what suggestions you receive, and how things are shown on the screen, all based on your personality.

This research is important for decision-makers, steering committees, commentators, and large companies to manage their accounts on social media platforms, and understand and analyze things that concern emotions and feelings.

1.8 Dessertation Structure

The following chapters, beginning with chapter two, outline the dissertation's structure after the introductory chapter:

Chapter Two: Literature Review: This chapter illustrates past research and publications relating to the study's objective: Social Media Platforms, The Big Five Personalities, The Big Five Personalities Traits (Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism), and Predicting Personality From Social Media Platforms.

Chapter Three: Research Methodology: The methodology of this study is based on a mixture of quantitative and qualitative information. This combination offers more methods and

concepts that are specifically relevant to creating a system for managing inventory, thus this chapter describes how the study was carried out and how the data was gathered and explained.

Chapter Four: Results and Discussion: This chapter includes an explanation and interpretation of the findings and results and will put the theoretical framework's theories into practice by demonstrating how the researcher used the methodology to conclude the results of this research.

Chapter Five: Conclusions, and Future Work: Accordingly, a discussion of findings, conclusions, and suggestions for future research work will be highlighted. In addition, this chapter will conclude the research's successes and present future enhancement to improve the results.

Chapter Two: Literature Review

2.1 Introduction

The purpose of this study is to address the connection between users' personality situations and the posts they publish on the X platform (Twitter). The research will utilize suitable data science methods to classify individuals' personalities based on an analysis of their tweets. The goal of the study is to determine how people's online expressions correspond to their underlying personality by looking at the content of their tweets and using relevant personality categorization data science methods and algorithms.

With the advancement of technologies nowadays, studies on social media analysis have become significant in data science knowledge (Dwivedi et al., 2021). It is possible to quantify the expressions, sentiments, and emotions that individuals publish on Twitter. These postings frequently represent certain emotions or originate from particular psychological scenarios. Therefore, using text analysis or machine learning techniques to ascertain someone's emotions and moods on platforms such as X Platform (Twitter) became possible (Maheshkar & Sarin, 2022). Scientists analyzing online social network platforms can find excellent environments for data science and data analysis, with applications including marketing, sales, and political understanding. Social media analysis is a popular subject in modern science. (Felt, 2016).

Few prior studies have examined the relationship between personality traits and online networks in greater detail. X (Twitter) platform data analysis for personality characteristic predictions is a relatively new field of research (Dhall & Hoey, 2016). Researchers have taken two main routes to explore the personality traits of users on social networks, the initial method entails employing diverse machine learning techniques to devise models solely grounded in online social actions, and the alternative approach broadens the scope by incorporating language cues related to personality traits (Oberlander & Nowson, 2006). For instance, Rao et

al. (2008) emphasized the importance of recognizing individual personality efficiently and reliably. They proposed a method for Big Five personality recognition from social media data, enabling the analysis of social media content and training models on the dataset. This approach can visually depict the character of social media users through radar plots, facilitating the identification of anti-social behaviors. Additionally, this method can serve as a hub module in online shopping and matrimonial sites and assist recruiters in hiring candidates. The costeffectiveness and practicality of this approach make it a valuable tool in various applications.

2.2 Social Media Platforms

The realm of social media has witnessed an annual expansion of 5.1%, attracting over 227 million fresh users in 2022. Research reveals that the total count of social media users for 2022 reached about 4.7 billion, making up approximately 59 percent of the world's population. Remarkably, 9 out of every 10 internet users incorporate social media platforms into their daily online routines. Nonetheless, it's very important to acknowledge the presence of fraudulent and duplicated accounts in these figures. On an average basis, individuals visit around 7.5 distinct social media platforms each month and spend roughly 2.5 hours every day on these platforms (Simon Kemp, 2023).

X Platform remains one of the largest social media platforms, with seven among them boasting more than one billion active daily users. Meta owns four of these platforms. It's noteworthy that this data underscores that 36.8 percent of the global population is part of the X Platform user base, solidifying its position as one of the most widely embraced social media platforms (Simon Kemp, 2023).

People engage with social media platforms primarily to stay connected with near and far friends and family, make use of comfort time, stay updated with news and narratives, uncover a range

of content like articles and videos, explore what's currently popular, look for ideas for activities and shopping, discover products to buy, take part in conversations and express their viewpoints, watch live broadcasts, establish new relationships, keep up with sports happenings, join communities that share their interests, and access content from beloved brands (Simon Kemp, 2021).

As of January 2021, the population of Palestine was 5.16 million. The number of internet users in Palestine was 3.65 million, indicating an internet usage rate of approximately 70.6% in January 2021. The total number of users on social media platforms in Palestine was 3.1 million. Notably, there was an increase of about 400 thousand users between 2021 and 2022, representing a growth rate of +15% (Source: Simon Kemp, 2021).

2.3 The Big Five Personalities

The idea of the five fundamental dimensions of personality has been widely investigated through various research studies over time, and the proof supporting this theory has considerably advanced. These five personality traits are well known as ***Openness***, ***Conscientiousness***, ***Extroversion***, ***agreeableness***, and ***neuroticism***.

Scholars have been trying to use psychological characteristics to study people's behavior before the Big Five fundamental personality traits notion was developed. Gordon Allport once gathered over 4,000 attributes; these were subsequently narrowed down to 16 (Kellan M, 2017). Nonetheless, this method's intricacy was still viewed as a drawback, leading to the development of the five fundamental characteristics of personality as a briefer and generally acknowledged structure (Najm Al-Zaytoonah, 2019).

2.4 The Big Five Personality Traits:

2.4.1. Openness

Individuals who exhibit high levels of openness usually have a diverse array of interests and tend to approach decision-making with a sense of adventure. They excel in exploring new territories and often display creative thinking. Conversely, those with lower levels of openness might find themselves confined to conventional thinking patterns, potentially impeding their problem-solving capabilities when faced with unfamiliar situations (Tarka et al., 2022).

2.4.2 Conscientiousness

A trait commonly found in finance and science professionals is good impulse control, coupled with the ability to exhibit goal-directed behavior. These individuals are known for their structured and organized approach to work, effectively carrying out their duties. They excel in planning and analyzing their behavior to understand its impact on others. These traits are particularly beneficial for HR personnel and management levels within an organization, as they help balance various roles and responsibilities. An example of this personality type is someone who consistently plans for meetings and regularly checks in with others to ensure proper organization around specific events. On the other hand, individuals who are low in conscientiousness tend to avoid the proper execution of tasks and are prone to procrastination, particularly with important projects (Boulianne, 2015; Carducci et al., 2018).

2.4.3 Extroversion

A person who feels excited to be with others and enjoys social interactions is commonly known as talkative and highly expressive emotionally. These individuals often thrive on being the center of attention and tend to have a large circle of friends (Shipankar et al., 2022).

Conversely, the opposite of this personality type is introversion. Introverts prefer solitude and may feel drained or have less energy in social situations. Extroverts, on the other hand, are

typically more public-facing and often hold roles in various fields such as politics, sales, and marketing. They are frequently seen as natural leaders.(Skues et al., 2012).

2.4.4. Agreeableness

People with high degrees of agreeableness tend to demonstrate kindness, affection, and trust toward others. They are eager to help and display highly prosocial behavior, being willing to share, cooperate, and provide comfort to those in need (Gallo et al., 2020).

Additionally, agreeable individuals possess the ability to empathize with others, which further contributes to their agreeable nature. In contrast, individuals who exhibit disagreeableness manifest in socially unpleasant behavior, often showing negative feelings towards others and displaying a lack of concern or caring. They may not take other people's problems seriously (Markovikj et al., 2013a).

2.4.5. Neuroticism

A person who is emotionally unstable, experiences sadness, and is often moody may suffer from a psychological disorder. People with this condition react to stress and situations they see as threats in their everyday lives with increased physical and emotional responses. They may frequently experience mood swings, irritability, and anxiety (Bai et al., 2012a).

The main trait associated with neuroticism is anxiety, which causes individuals to be prone to overthinking and perceiving risks and stressful situations. On the contrary, Individuals who possess low levels of neuroticism typically exhibit a steadier and more flexible mindset when facing stressful circumstances. They are less likely to become overwhelmed by the thought of potential stress-related factors (Boulianne, 2015).

2.5 Previous Studies

Over the past decade, several studies have focused on exploring the connection between personality traits and the usage of social networks. Personality traits have also been linked with the most widely used social network system in China (Bai et al., 2012b). Among the methods employed, Decision Trees demonstrated the most favorable results, achieving accuracy rates of 69-72 percent by considering factors related to users' network engagement and emotive linguistic cues extracted from their status updates and blog posts.

Tan et al. (2011) adopted two regression techniques, specifically m5sup/Rules and Gaussian Processes, to construct predictive models for personality. They took into account users' social media data and analyzed elements such as structural characteristics, personal information, activities, and preferences, alongside linguistic traits obtained from users' statuses using the LIWC tool. It is worth noting that both studies faced limitations in their participant sampling, which affected the generalizability of the results. One study was limited to the Chinese population, while the other focused on the authors' Facebook friends, leading to a lack of demographic diversity in the samples.

Similarly, Markovikj et al. (2013b) Examined the importance of extracting information from Facebook data for predictive personality modeling. Their study investigated whether it's possible to create models that can predict user personalities using a defined set of characteristics derived from Facebook data. They also analyzed how well various classification methods worked for this task, and found that restricting the features used for classification to those that strongly correlated with personality traits improved the predictive modeling performance. The initial research results were encouraging, indicating that selecting the most indicative features can enhance classifier accuracy. Gaining qualitative insights from the massive pool of available

information is just the initial step toward seeking meaningful and plausible explanations for the correlations between personality traits and user interactions within social networks.

Few studies have utilized a significantly larger number of instances from collections of data used to investigate connections between personality traits and the activities conducted on social media (Bachrach et al., 2012). For example, Kosinski et al. (2013) investigated the links between individual traits and Facebook “Likes”. However, these studies did not focus on the rich linguistic aspects present in language usage on social networks.

In a more recent study, Sayeb et al. (2021) achieved an impressive 87.5% accuracy using the Radial Bases Function (RBF) classifier based on Facebook account data. Their research predicts certain personality traits based on the interaction between Facebook users.

In the meanwhile, Ortigosa et al. (2014) developed a web app called TP2010 that collects users' personality traits and uses machine-learning techniques to analyze the data.

Another study by Schetgen et al. (2021) revealed that Facebook pages and categories were the most important data types, with age, education, residence, and interests being significant predictors.

Skues et al., (2012) focused on the relationship between narcissism, openness, extraversion, and self-esteem, finding higher levels of openness linked to increased Facebook use. The research also indicated that increased feelings of loneliness were connected to having a larger number of friends on Facebook.

Liu & Campbell (2018) investigated the prediction of social networking footprints for personality characteristics and found that digital footprints match with behavior's normal predictive higher bound.

Similarly, Albayati & Altamimi (2019a) carried out a study utilizing supervised and unsupervised machine learning methods in addition to different attributes and labels to distinguish between real and fake profiles. In their research, the authors used two unsupervised algorithms namely k-Means clustering and k-Medoids clustering, three supervised algorithms namely k-Nearest Neighbors, Support Vector Machine, and ID3 decision tree, and two supervised methods namely k-Nearest Neighbors (K-NN), and support vector machine (SVM). Additionally, the authors used the k-NN algorithm for data cleaning and to handle missing values in the data. The study concluded that the supervised algorithms, which were trained on cleaner data provided more accurate results compared to the unsupervised algorithms. However, the study also has some drawbacks. For instance, the study did not explore additional techniques for cleaning data in the context of unsupervised algorithms, which could have improved accuracy. Additionally, the inclusion of more algorithms could have further enhanced the results. Moreover, the study did not consider personality mining concerning the available data which may be limiting its scope.

In a second study, Albayati & Altamimi (2019b) analyzed the natural vocabulary used in online content to forecast users' types of personalities using the Myers-Briggs personality type system (MBTI). The MBTI is a personality system that reduces a wide variety of people's behaviors, views, habits, and traits into a few distinct and understandable psychological profiles. It offers perceptive data that might enhance one's comprehension of others as well as oneself.

The study utilized text mining from social networking platforms like X Platform as well as Instagram to forecast MBTI personality types. The work used neural network approaches to determine the personality categories of the writer and classify them based on their verbal qualities. The authors concluded that people display a variety of actions that align with their

personalities. Accordingly, the authors were able to divide people into 16 distinctive personalities grouped into four main groups. The study lays the groundwork for potential advances in the field of psychology. Ultimately, the proposed approach may find application in customized social media relationships market analysis, and psychological assessment.

Bokishev (2022) used a Kaggle gathered dataset and compared it against the number of the majority of the different individuals included in their dataset in order to assess the global bias in tweets based on the number of words per tweet. Depending on a variety of parameters, the multinomial logistic regression approach is a helpful instrument for predicting classified placing or evaluating the probability of falling into a particular category. It makes adding more than two categories to a dependent factor easier. The precision attained after combining the training information and the model was 72% for training data and 63% for testing data, demonstrating an adequate distinction to address overfitting. Support Vector Classifiers (SVM) are powerful techniques used for classification and regression tasks. They have shown promising results in various fields, including natural language processing. After training the model on the dataset, it achieved 82% accuracy for training data and 66% accuracy for testing data. However, there is a significant difference between test and training performance, indicating the presence of overfitting. XGBoost, a gradient-boosted decision tree technique, is commonly used for classification, regression, and ranking tasks. It offers a great balance of prediction performance and processing time, making it well-suited for NLP problems. After fitting the model with XGBoost, the training accuracy achieved was 92%, while the testing accuracy was 67%.

As a result, this research uses XGBoost to evaluate the results. And they also used some posts on Instagram for their research dataset. The weaknesses are the comparison of all models which shows a common problem of high difference between training and testing accuracy, indicating overfitting. To address this, increasing the dataset size could improve performance. Additional

hyperparameter tuning did not significantly improve results. The overfitting issue limits the practical usage of the proposed method

Celli & Lepri (2019a) discussed two popular psychological tests for assessing personality the Big Five and the Myers-Briggs Type Indicator of MBTI. The Big Five model uses five scales to define personality traits, while the MBTI uses four binary classes to create 16 personality types. Correlation analyses show connections between these measures. The MBTI's popularity stems from its easier data collection. In their paper, Celli & Lepri (2019a)aimed to ascertain whether utilizing machine learning makes it more straightforward to predict classes related to either the Big Five personality traits or the MBTI personality types. Two X Platform datasets are collected in multiple languages and annotated with Big Five and MBTI types. This study contributes to understanding the heuristic power of these models in machine learning tasks.

“The MBTI model formalizes personality types as classes, while Big Five as scores.” Personality computing from text originated over a decade ago, initially identifying Big Five traits from users’ blogs and self-presentations. Currently, related fields emerged, exploring personality computing from diverse signals like recorded meetings. Research on the MyersBriggs Type Indicator (MBTI) was limited until around 2015, while the Big Five gained traction for automatic personality prediction from various sources, notably Twitter. Text features like sentiment, Part of Speech (PoS) tags, and psycholinguistic tags, along with metadata such as followers, network density, and likes, were commonly used. The collaboration between computer science and personality psychology has led to valuable insights, demonstrating strong correlations between personality types and attributes like education, religion, and political preferences extracted from social media profiles. Additionally, correlations were found between personality types and user popularity and influence on social media.

Celli & Lepri (2019)) collected two X Platform datasets, each with 900 users, labeled for MBTI and Big Five personality types. For the Big Five, researchers used advanced search to get diverse test results. Both datasets covered English, Italian, Spanish, and Dutch languages. The MBTI set showed more tweets than the Big Five. Users' tweets were concatenated and limited to 40 per user for balance. The resulting datasets had 265K and 290K words, respectively. Big Five classes were balanced, while MBTI had imbalances. Experiments aimed to balance classes, test various feature combinations, and assess machine learning algorithms' performance in predicting personality types from both models. researchers compared the performance of algorithms in predicting the study conducts nine binary classification tasks to distinguish between personality classes associated with both the Big Five and MBTI frameworks. Celli & Lepri, (2019b) utilized three types of features:

- Character n-grams (1000 features): Extracted 1000 character bi-grams and tri-grams from tweets.

- LIWC match ratio (68 features): With this statistic, you may figure out how many words in the LIWC dictionaries match each other.

- Metadata (10 features): Included various profile-related ratios and attributes.

They also applied a subset selection algorithm for feature selection and balanced classes and employed SVMs along with a meta-classifier. The evaluation used 10-fold cross-validation, with accuracy and average metrics reported. researchers also presented averages for the four Big Five traits correlated with MBTI for maximum comparability. But they didn’t delete the English stop words and that will cause a big bias and bad results. The research concluded that SVMs generally perform better in predicting MBTI classes compared to the Big Five, with more consistent predictions for MBTI. On the other hand, SVMs with specific kernels show varying success in predicting Emotional Stability and Agreeableness, suggesting non-linear

separability. MBTI predictions appear more stable due to class balancing, LIWC usage, and feature subset selection. Specific features, such as references to others or first-person pronouns, hold predictive power for certain traits. An English-only experiment minimally impacts results, showing that English is predominant. So, language-specific best features shift, with hashtags for Extraversion in Big Five and anger for Extraversion in MBTI It has been discovered that words associated with emotions in the English language significantly influence the attribute of agreeableness. The researchers divided the Big Five scores into predetermined categories to compare the two personality models. Compared to the MBTI, this categorization made predicting Big Five classes more difficult. However, specific traits within the Big Five exhibited enhanced performance when non-linear models were employed. A limitation arises from the diverse sets of users present in the datasets, which could potentially introduce biases in the classification process. Unfortunately, due to the significant effort needed for thorough data collection, collecting data for the same persons labeled with both MBTI and Big Five qualities using X Platform inquiries is not practical (Celli & Lepri, 2019).

For their recruitment study, Faliagka et al., (2012a) proposed a method where all applicants have to share their blog and a LinkedIn account, thus 100 applicants who met these requirements were assembled. Using the Google blog search API, a random selection of these applications was made. The selection criteria, which were clear from blog metadata and LinkedIn profiles, focused on a technological background. To locate blogs in the "technology" niche, the Google profile search API was used. Only bloggers with linked LinkedIn accounts were eventually included after manual examination. Technical blogs and other blogs lacking autobiographical content were removed because they couldn't reveal anything about the personalities of their creators. The first 100 blogs that the profile search API retrieved and satisfied the requirements for the study were chosen as the applicant corpus. This selection process maintains independence from candidate selection standards, ensuring unbiased study

results. Furthermore, information about three other professional career jobs with different requirements was gathered by investigators from an unidentified IT business. This unique method evaluates the system's capacity to match profiles of applicants with appropriate job openings. For instance, extraversion is valued in the sales engineering function, expertise is given priority by senior developers, and trust and knowledge are used to assess new developers. Despite the study's focus on technical roles, various other sectors can also benefit from the techniques employed. Scientists used an innovative technique in the present research to forecast how hiring managers will evaluate a candidate's fitness for a particular post. The investigation also contributed to a better understanding of automatic candidate screening in online recruiting systems. In contrast to the old method of depending just on resumes, it explicitly stressed the use of individuals' LinkedIn accounts. The research emphasized that successful candidate ranking requires access to comprehensive profiles and recruiters' criteria. Through careful parameterization, including assigning weights to selection criteria, the approach consistently generated candidate rankings aligned with expert recruiters' assessments. Additionally, the study demonstrated the feasibility of inferring personality traits from candidates' blogs, offering potential selection criteria for positions valuing such traits. “This allows employers to access the applicant’s full employment history and group of contacts, and gives them the chance to automatically evaluate the candidate’s profile for a broad range of job positions without the complexity of parsing a full-text resume” (Faliagka et al., 2012b).

Through the internet social networking networks and related innovations are widely used in modern society today. This has sparked a keen interest in learning about user behavior across various social media platforms. In their research, Al Maruf et al., (2015a) explored whether people act the same way on different platforms. They look at how people use two big online platforms – X Platform and Disqus – to better understand a person as a whole. By studying how people write and the traits they show on both platforms, they find connections between

personality traits and how people use X Platform and Disqus. The results also show that social media can change what subjects and topics a person talks about. Both platforms let people share their thoughts on many things, but X Platform and Disqus have different ways of figuring out how people feel. In particular, Disqus is better at figuring out how someone feels about a specific thing. By looking at what people do on both platforms, researchers learn a lot about what people are interested in and what matters to them. This shows that people use different social networks for many widely different reasons, so it's not enough to understand someone from just one platform. In this study, they look at the profiles of 105 people on both Disqus and Twitter. Disqus comments consistently offer a clear portrayal of user sentiment on these subjects. This study reveals that in more than 90% of the cases, Disqus posts distinctly convey the sentiment polarity. It's noteworthy that tweets frequently contain neutral sentiments, surpassing the count observed in Disqus comments.

Disqus emerged as more powerful in identifying sentiment compared to X Platform(Al Maruf et al., 2015b). While tweets occasionally displayed sentiment ambiguity or neutrality, Disqus comments often provided clear indications of user sentiment on topics or entities. Over 90% of cases demonstrated clear sentiment polarity in Disqus comments. Additionally, finding that people's topics of discussion varied between the two platforms. Common topics were fewer in number across both platforms, and many users engaged exclusively on distinct topics in either X Platform or Disqus. The research underscores the influence of the platform itself on a person's subjects of discussion, shedding light on the multifaceted nature of online identities across different social networks. (Al Maruf et al., 2015b).

De Montjoye et al., (2013a) conducted research for mining personality traits by mobile use. The study implied that the way people use their mobile phones can reveal insights into their personalities, and this idea has gained significant attention as mobile subscriptions have reached 6 billion worldwide. Mobile carriers provide phone usage data to researchers and

commercial partners, potentially offering an unobtrusive and cost-effective way to gauge personality traits. This information could be useful for tailored marketing and personalized services. Recent advancements in data collection and machine learning have shown that psychological states and traits can be inferred from digital technology use. For instance, patterns in social media usage or mobile app usage can predict personalities. However, existing methods often require extensive information or specific tracking apps, limiting their applicability for large-scale studies. This study attempts to show that basic data that is available to all mobile phones and telecommuting service providers may be used to properly determine user personalities. The study introduces measures based on common phone logs that are psychology-informed, including Basic phone use, Active user behaviors, Mobility, Regularity, and Diversity. Extroversion, agreeableness, conscientiousness, openness to experience, and emotional stability are among the attributes that can be predicted by these measurements. The study investigated several predictors for the Big Five personality traits. All of which were significantly related to personality, an SVM classifier was built. The model's accuracy in predicting levels of neuroticism, extraversion, conscientiousness, agreeableness, and openness was significantly higher than the baselines at 54%, 61%, 51%, and 49%, respectively. The prediction accuracy of neuroticism increased to 63% when gender was included as a predictor, which is consistent with earlier research linking neuroticism and gender. The entropy of contacts for extraversion and agreeableness, as well as mobility-related metrics (distance traveled and entropy of places) for neuroticism, were notable markers for prediction. and variance in call times for conscientiousness. These indicators matched previous research suggesting associations between these traits and aspects of social behavior, organization, and punctuality.

De Montjoye et al. (2013b) distributed Android smartphones with the Funf open-sensing framework to a group of participants. The participants' phones were used to gather a variety of

behavioral data using this methodology. However, the study only examined information found in the logs of common carriers, such as phone calls and text messages. These Call Data Records have become increasingly popular for computational social science research. After excluding participants with fewer than 300 calls or texts per year and those who didn't complete personality assessments, the final sample consisted of 69 participants. The participants were 51% male, with a standard deviation of 6.1 years and an age distribution of 30.4 years on average.

EVRİM & NİSSOUL (2019a) conducted a study of social media personality detection benefits. In their study, they explored the connections between personality traits, life satisfaction, and 12 distinct interest groups of users based on data from myPersonality Facebook dataset. The research derived relevant principles of association and identified connections between various user attributes. The study utilized the Apriorist algorithm within the open-source tool called WEKA (Waikato Environment for Knowledge Analysis) to
produce useful association rules from a dataset. In this project, three specific WEKA filters are employed for balancing: SMOTE, Spread Subsampling, and Resampling. These filters offer different approaches to achieve balance in the data, with each class being balanced to specific itemset sizes. The paper analyzed interest categories using 8 personal characteristics, and the study experiments revealed that there are additional user attributes (e.g., health, childhood experiences) that can serve as markers of different types of interest. With more comprehensive data, user interests can be examined in a wider context, taking larger sets of grouped interests into account.

Stillwell et al. (2014a) explored the valuable insights from user-generated material on social media sites, particularly focusing on emotions and personality characteristics without relying on traditional questionnaires by analyzing nearly 1 million Facebook status updates. The study investigated the connections between emotions expressed in these posts, the users' age, gender,

personality, and the timing of the posts. The findings of this article reveal several interesting patterns. Female social media users tend to provide more emotions in their posts of status versus male users. Moreover, there is a relationship between age and emotional sharing, with older Facebook users being more prone to share their feelings compared to younger users. The analysis of seasons shows that people post about emotions less frequently during summer, While December appears to be the month when people are most likely to talk positively to their friends.

Stillwell et al. (2014b) investigated the connection between user personality traits and their posts. The study shows that people who have an open personality prefer to express their emotions more openly, whereas people who have a neurotic personality are more reticent to communicate their emotions.

In this study, data from the myPersonality project served as the foundation for this study's methodology and conclusions. Users completed a typical personality questionnaire and were permitted to record their answers and Facebook accounts. A user may not have access to all of the data in the data set, which also includes information on demographics, friendships, Facebook activity, status updates, and Big Five Personality Scores. 5,865 users with comprehensive information on age, gender, personality assessments, and status updates were the subject of the research. These users' 969,035 status updates are included in the data set. Using predetermined score thresholds, the researchers divided the users into personality-based categories. Each user in the data set has a personality score for each personality feature, which ranges from 1 to 5. Users were split into two groups by the researchers in order to facilitate analysis: those who overtly display a particular trait and those who do not. The threshold values and the number of users falling into each trait category are shown in Table 2.1. They defined these groups using the same thresholds as a previous data set. In one study, for instance, a user is classified as an "extrovert" if their Extroversion score is at least 3.60, and using this criterion,

researchers discovered 2,971 such individuals in the data set. It's important to note that this approach of categorizing users into binary groups based on the five personality dimensions is a somewhat simplistic method. A more nuanced analysis considering the gradual spectrum from Introversion to Extroversion could yield additional valuable insights.

Table 2.1. The threshold score for each personality-based category (Stillwell et al., 2014b)

Personality	Threshold	Number of users
Extroversion	3.60	2,971
Openness	3.80	3,284
Agreeableness	3.55	3,110
Conscientiousness	3.50	3,071
Neuroticism	2.80	2,631

The findings demonstrated a relationship between user traits and emotional reactions. In comparison to neurotic users, female users and older users frequently show more emotions in their status updates. Open users are more willing to express their opinions on a range of topics. The analysis of timestamp data revealed that emotional status updates are more frequently posted on workdays than on weekends and that they are most frequent in December and least frequent in the summer.

In conclusion, significant correlations were found between the selected features (age, gender, personality) and users' emotions. The researchers plan to develop a model for predicting the most likely upcoming emotion for each user, considering factors like time, demographics, and personality. This predictive model could be valuable for personalized services, enabling targeted content delivery based on users' emotions. (Stillwell et al., 2014b).

“Personality is a word that comes from the Latin word persona, which refers to the mask used by the actors in a theater”. This idea comes from the idea that a person's personality is their external features. Adding to this, (Mairesse et al., 2007) define personality as a group of traits

that collectively make up who a person is, including their conduct, temperament, emotions, and mentality. However, this collection of traits can be quite large, making it difficult to adequately sum up either the user's own or another person's personality.

The Big Five personality test will be administered to groups of individuals as the next research challenge in order to build a repository for training categorization algorithms (Lima & De Castro, 2013). This will make it easier to create new connections between traits and personality qualities. In addition, testing additional categorization techniques, such as Support Vector Machines (SVM) is planned to be able to predict the personality qualities of a Facebook user, researchers applied machine learning approaches (Support Vector Machine, Nearest Neighbor, and Naive Bayes) (Lima & De Castro, 2013). The user's social network attributes, the information in their status updates, and the frequency of those updates were all examined by researchers. They also took into account individual traits to improve categorization outcomes. For instance, when concentrating only on the "anger" LIWC (Linguistic Inquiry and Word Count) function, the results for recognizing emotional stability were better than any other approach used for this specific personality trait (Farnadi et al., 2013).

Gou et al. (2013a) explored how personality affects how users make decisions, communicate, and respond to different circumstances. In this paper, the Myers-Briggs Type Indicator (MBTI) personality classification system is briefly introduced. The planning of an experiment run on the Apache Spark platform then becomes the study's main focus. The paper provides a program called "PersonalityViz" that was created to make it easier for people to understand the personality traits that their social media activity, specifically their tweets, reveal about them. To determine a person's Big Five personality traits based on their tweets, the system uses the Linguistic Inquiry and Word Count (LIWC) text analysis tool in conjunction with LIWC/Big Five personality correlations. Users of the program can explore and study their personality features over time using an interactive visual interface. In order to help users comprehend how

these characteristics are determined from the pertinent tweets, it also provides visual evidence (Gou et al., 2013b).

In order to increase the precision of sentiment analysis on X Platform data, Samih et al. (2022) offered a novel method named "Improved Word Vectors for X Platform Sentiment Analysis" (IWVTSA). The focus of the study is on sentiment indicators discovered through lexical and syntactic analysis. The research suggests an alternative strategy by combining the "doc2vec" and the "tweet2vec" word embedding methods to produce a full set of sentiment features for tweets. These word embeddings and word sentiment polarity ratings are combined. This approach could serve as a strong foundation for creating sentiment analysis tools for social media. In the future, researchers aim to enhance the approach by using more advanced techniques and additional information like emojis or hashtags. Researchers are also interested in applying sentiment analysis for recommendation systems, given the growing importance of user reviews in various fields.

Quercia et al. (2011a) revealed two key findings. First, different types of X Platform users share both common personality traits and distinctive differences Users, regardless of their category (listeners, popular, highly-read, and influential), generally exhibit emotional stability (low Neuroticism) and many shows extroverted tendencies. While these ideas have been informally recognized, they've been challenging to precisely quantify. Notably, popular users seem to be more creative, whereas influential users tend to be more organized.

The ability to anticipate a user's personality using publicly available data is the second important realization (Quercia et al., 2011b). This finding points to interesting future paths in several fields:

1. Marketing: By understanding the relationship between customer personality and marketing tactics, one might adapt advertisements to match users' expected preferences, increasing their receptivity.

2. User Interface Design: In addition to content, it is also possible to match a social media platform's general aesthetic to particular personality qualities. This idea, referred to as "website morphing," might be used.

3. Recommender Systems: Taking into account user personality could improve the predictability of music recommender systems, given the well-established link between personality and musical preference.(Quercia et al., 2011b).

To ensure the predictive power of the model, researchers applied a standard maximum likelihood approach to estimate the parameters, using training data and then tested the model on separate testing data. The research delved into predicting extraversion using regression models, highlighting specific linguistic features and mood differences between extroverts and introverts. The results align with previous observations, emphasizing the distinctive language patterns and emotional expressions exhibited by these personality types (Nguyen et al., 2011).

Quercia et al. (2012a) analyzed a sample of 172,952 Facebook users from the United States. These users had taken the Big Five personality test, had between 30 and 1000 social contacts, and ages ranging from 18 to 54 The group consisted of 60% women and 40% men, with a median age of 23. The research consisted of three main steps:

1. Making Useful Predictions for Facebook Contacts: It investigated how personality characteristics and the logarithm of the number of social connections relate to one another. There were some weak associations, with Extraversion and Age having some effect. Older individuals generally had fewer Facebook contacts. Neuroticism had a

very weak negative correlation. These findings aligned with prior research about both offline and online popularity.

2. Statistical Significance of Predictors: A regression model was constructed to predict the logarithm of the number of contacts based on personality traits, sex, and age. Extraversion and Age had the highest statistical significance in predicting the number of Facebook contacts. Adding other variables did not significantly improve the model's fit.

3. Nature of Linear Relationships: While Extraversion was found to correlate with the logarithm of Facebook contacts, the strength of the linear relationship wasn't clear. The study segmented users into groups based on their social contacts and analyzed the average personality scores for each group. The results showed that Extraversion and the logarithm of contacts had a distinct and significant linear connection. It was predicted that extreme extroverts would make twice as many contacts as extreme introverts.

In summary, in this study, extraversion's impact on Facebook popularity was specifically explored with respect to personality attributes. The research showed that Extraversion and Age significantly predict the number of Facebook contacts, and the linear relationship between Extraversion and contacts was consistently evident across user groups.(Quercia et al., 2012b).

In the age of Big Data in telecommunications, effectively allocating resources has become a key challenge. Chen et al. (2019) investigated a way to allocate broadband inside an identical statistical structure, taking into account users' characteristics and how they use their smartphones. Companies may maximize their limited resources and obtain useful insights by customizing broadband distribution to consumers' personalities while taking into account various media circumstances. This is accomplished by predicting people's information

consumption based on an analysis of their psychological attributes and then allocating broadband appropriately. The suggested method improves customer service, makes optimal use of connectivity, and adjusts to channel circumstances in real-time.

Internet service companies may more effectively distribute cellphone broadband by using both diagnostic and predictive judgments. This allows them to better match the personality traits of their customers and maximize the utilization of resources. Through the prediction of client demographics and actions, this strategy also promotes corporate insights. Under fading network circumstances, the proposed distribution of bandwidth approach accounts for breakdown capability and likelihood. Greater capacity and better service quality are enjoyed by users who are more likely to need more connectivity, while serious withering conditions can be resolved by allocating additional broadband.

2.6 Summary of previous research

Table 2,2 summarizes the previous research as discussed in section 2.5.

Table 2,2. Summary of previous research

Title	Methods Utilized
Tan et al. (2011)	Adopted two regression techniques, specifically m5sup/Rules and Gaussian Processes, to construct predictive models for personality.
Markovikj et al. (2013b)	Examined the importance of extracting information from Facebook data for the purpose of predictive personality modeling.
Kosinski et al. (2013)	Investigated the links between individual traits and the Facebook “Likes”.
Sayeb et al. (2021)	Their research predicts certain personality traits based on the interaction between Facebook users.
Ortigosa et al. (2014)	Developed a web app called TP2010 that collects users' personality traits and uses machine learning techniques to analyze the data.
Schetgen et al. (2021)	Revealed that Facebook pages and categories were the most important data types, with age, education, residence, and interests being significant predictors.
Skues et al., (2012)	The relationship between narcissism, openness, extraversion, and selfesteem, finding higher levels of openness linked to increased Facebook use.

	The research also indicated that increased feelings of loneliness were connected to having a larger number of friends on Facebook.
Liu & Campbell (2018)	The ability of social media footprints to predict personality traits, and discovered that the data is consistent with behavior's typical predictive upper limit for forecasting personality.
(Albayati & Altamimi, 2019a)	Utilizing supervised and unsupervised machine learning methods. To evaluate the presented data, the researchers utilized different attributes and labels to distinguish between real and fake profiles.
Bokishev (2022)	Shows the count of most of the different personalities shown in their dataset to check the bias of the world in the tweets based on the word count per tweet.
Celli & Lepri (2019a)	Discussed two popular psychological tests for assessing personality the Big Five and the Myers-Briggs Type Indicator of MBTI. The Big Five model uses five scales to define personality traits, while the MBTI uses four binary classes to create 16 personality types. Correlation analyses show connections between these measures.
Faliagka et al., (2012a)	Researchers introduced a novel method to predict how human recruiters assess the suitability of job applicants for specific positions.
Al Maruf et al., (2015a)	Studied how people act on two different platforms X Platform and Discus. By studying how people write and the traits they show on both platforms, they find connections between personality traits and how people use X Platform and Disqus.
De Montjoye et al., (2013a)	Research for mining personality traits by mobile use. The study implied that the way people use their mobile phones can reveal insights into their personalities.
EVRİM & NİSSOUL (2019a)	Conducted a study of social media personality detection benefits. Research has shown that extracting personality traits from social media texts can improve recommender systems, particularly in marketing.
Stillwell et al. (2014a)	Explored the valuable insights from user-generated material on social media sites, particularly focusing on emotions and personality characteristics. The study investigated the connections between emotions expressed in these posts, the users' age, gender, personality, and the timing of the posts. The findings of this article reveal several interesting patterns. Female social media users tend to provide more emotions in their posts of status versus male users. Moreover, there is a relationship between age and emotional sharing, with older Facebook users being more prone to share their feelings compared to younger users.
Lima & De Castro, (2013)	Predict the personality qualities of a Facebook user, researchers applied machine learning approaches

Gou et al. (2013a)	Explored how personality affects how users make decisions, communicate, and respond to different circumstances.
Quercia et al. (2011a)	Investigated how personality characteristics and the logarithm of the number of social connections relate to one another.
Chen et al. (2019)	Proposed an approach that enhances user experience, optimizes bandwidth usage, and adapts to real-time channel conditions. explores a method to distribute bandwidth based on users' smartphone habits and personality traits.

Chapter Three: Methodology

3.1 Overview

These sections provide an overview of the model developed for labeling and extracting personality traits from the dataset. Section 1 presents an overview of the dataset. Its characteristics and limitations. Section 2 provides an overview of the model used in the study, and the main steps used to achieve the objectives as set in Chapter 1. Sections 3 and 4 provide a detailed model overview, including the applied steps in detail, see Figure 3.1 for the methodology diagram.

3.2 Research Methodology

Figure 3.1 illustrates the methodology followed in this research. To summarize, the following steps were followed:

a) Data Gathering: In this research, the data is gathered from the Kaggle platform, a website that provides datasets and computational resources.

b) Data preparation: several steps and methodologies were used to prepare the data for analysis and modelling.

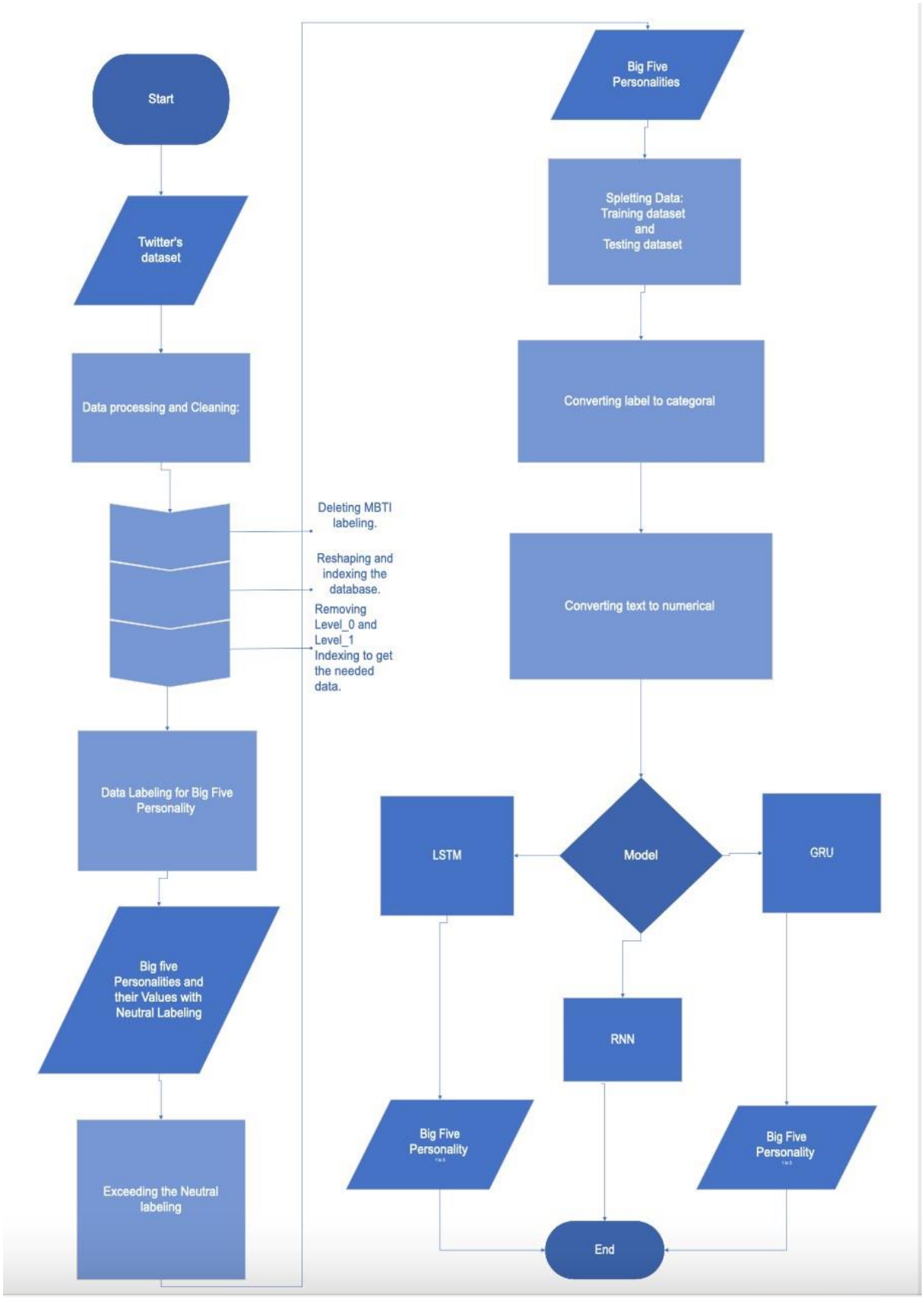


Figure 3.1. Research Methodology

3.3 Algorithms and Methods Utilized in this Work

In this research, several algorithms and methods were utilized, Table 3,3. Summarizes these algorithms and methods and how it was utilized.

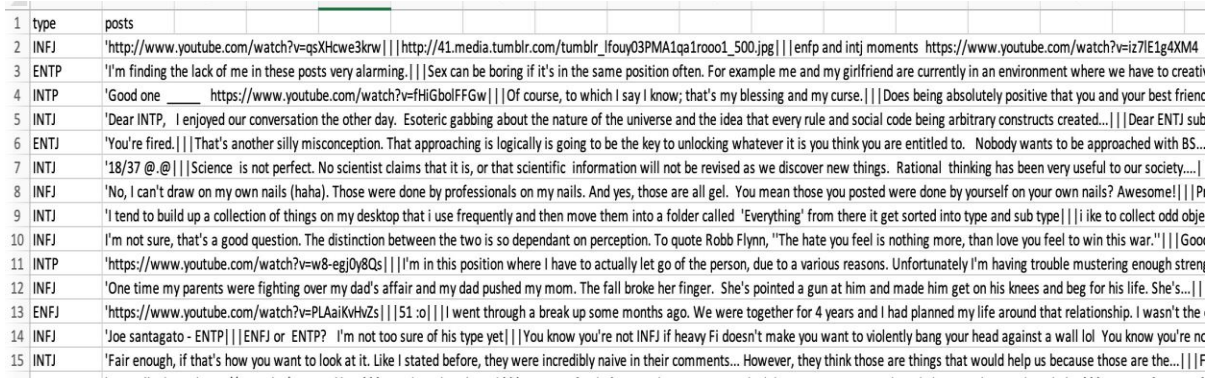
Table (3,3). Utilized algorithms and methods

algorithms and methods	Utilization Method
Word2Vec	word embeddings to extract features. This technique algorithm uses a natural network model to learn word associations from a large frame of text. Represents every distinct feature of a word with a particular list of numbers called vectors. Aims to capture semantic relationships between words by training on large text corpora
VADER	`sia = SentimentIntensityAnalyzer()`: This line creates an instance of the `SentimentIntensityAnalyzer` class from the NLTK library. This class provides a pre-trained model for performing sentiment analysis on text using the VADER sentiment analysis tool.
	The next part of the code defines a function named `label_sentiment(sentiment_scores)`. This function takes a dictionary of sentiment scores as an argument.

3.4 Dataset acquesition and preparation:-

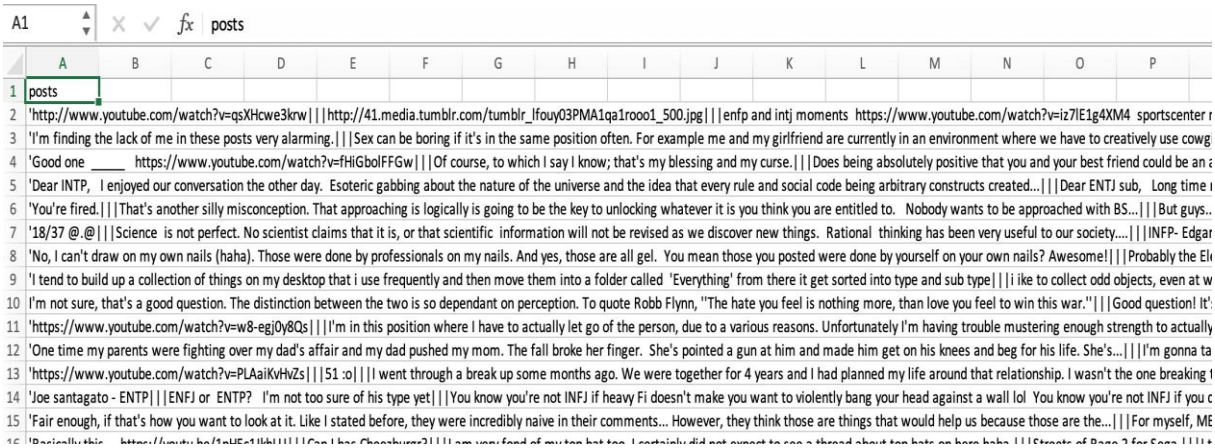
At the beginning of conducting this research, this study was about predicting Facebook users' personalities by analyzing the users' profile posts using machine learning techniques. The restrictions on Facebook APIs' prevented data extraction due to the privacy policy of the platform. As a result, there was no ability to find Facebook data over the web. The only Facebook data found provides only numbers and does not meet our goals. X Platform on the other hand is more flexible in providing users data. So, it was easy to find X Platform post datasets over the web. The dataset was pulled up from the Kaggle databases website. It was named "mbti_1.csv" and the size of the dataset was 25.6 MB. After unzipping it, the size of the dataset file became 62.8 MB. The X Platform dataset in its first look has 51 columns. The 1st column is for MBTI (Myers-Briggs Type Indicator) personality labeling for 8,675 records of X

Platform users, Figure 3.2 provides a view of the data included in the dataset. Each record represents a X Platform a person's account, while every person's account has 50 tweets. These records of tweets are the main feed for this study. The MBTI labeling in the dataset is deleted since it belongs to another personality mining technique that is not utilized in this study, keeping this labeling will affect the performance of the study.



	type	posts
2	INFJ	'http://www.youtube.com/watch?v=qsXHcwe3krw|||http://41.media.tumblr.com/tumblr_lfouy03PMA1qa1rooo1_500.jpg|||enfp and intj moments https://www.youtube.com/watch?v=iz7lE1g4XM4
3	ENTP	'I'm finding the lack of me in these posts very alarming.|||Sex can be boring if it's in the same position often. For example me and my girlfriend are currently in an environment where we have to creati
4	INTP	'Good one _____ https://www.youtube.com/watch?v=fHiGbolFFGw|||Of course, to which I say I know; that's my blessing and my curse.|||Does being absolutely positive that you and your best frienc
5	INTJ	'Dear INTP, I enjoyed our conversation the other day. Esoteric gabbing about the nature of the universe and the idea that every rule and social code being arbitrary constructs created...|||Dear ENTJ sub
6	ENTJ	'You're fired.|||That's another silly misconception. That approaching is logically is going to be the key to unlocking whatever it is you think you are entitled to. Nobody wants to be approached with BS...
7	INTJ	'18/37 @.@|||Science is not perfect. No scientist claims that it is, or that scientific information will not be revised as we discover new things. Rational thinking has been very useful to our society....|
8	INFJ	'No, I can't draw on my own nails (haha). Those were done by professionals on my nails. And yes, those are all gel. You mean those you posted were done by yourself on your own nails? Awesome!|||P
9	INTJ	'I tend to build up a collection of things on my desktop that i use frequently and then move them into a folder called 'Everything' from there it get sorted into type and sub type|||i ike to collect odd obje
10	INFJ	I'm not sure, that's a good question. The distinction between the two is so dependant on perception. To quote Robb Flynn, "The hate you feel is nothing more, than love you feel to win this war."|||Goo
11	INTP	'https://www.youtube.com/watch?v=w8-egj0y8Qs|||I'm in this position where I have to actually let go of the person, due to a various reasons. Unfortunately I'm having trouble mustering enough streng
12	INFJ	'One time my parents were fighting over my dad's affair and my dad pushed my mom. The fall broke her finger. She's pointed a gun at him and made him get on his knees and beg for his life. She's...||
13	ENFJ	'https://www.youtube.com/watch?v=PLAaiKvHvZs|||51 :o|||I went through a break up some months ago. We were together for 4 years and I had planned my life around that relationship. I wasn't the
14	INFJ	'Joe santagato - ENTP|||ENFJ or ENTP? I'm not too sure of his type yet|||You know you're not INFJ if heavy Fi doesn't make you want to violently bang your head against a wall lol You know you're nc
15	INTJ	'Fair enough, if that's how you want to look at it. Like I stated before, they were incredibly naive in their comments... However, they think those are things that would help us because those are the...|||F

Figure 3.2 The original dataset from Kaggle.



A1 fx posts

	A
1	posts
2	'http://www.youtube.com/watch?v=qsXHcwe3krw|||http://41.media.tumblr.com/tumblr_lfouy03PMA1qa1rooo1_500.jpg|||enfp and intj moments https://www.youtube.com/watch?v=iz7lE1g4XM4 sportscenter
3	'I'm finding the lack of me in these posts very alarming.|||Sex can be boring if it's in the same position often. For example me and my girlfriend are currently in an environment where we have to creatively use cowg
4	'Good one _____ https://www.youtube.com/watch?v=fHiGbolFFGw|||Of course, to which I say I know; that's my blessing and my curse.|||Does being absolutely positive that you and your best friend could be an a
5	'Dear INTP, I enjoyed our conversation the other day. Esoteric gabbing about the nature of the universe and the idea that every rule and social code being arbitrary constructs created...|||Dear ENTJ sub, Long time
6	'You're fired.|||That's another silly misconception. That approaching is logically is going to be the key to unlocking whatever it is you think you are entitled to. Nobody wants to be approached with BS...|||But guys..
7	'18/37 @.@|||Science is not perfect. No scientist claims that it is, or that scientific information will not be revised as we discover new things. Rational thinking has been very useful to our society....|||INFP- Edgar
8	'No, I can't draw on my own nails (haha). Those were done by professionals on my nails. And yes, those are all gel. You mean those you posted were done by yourself on your own nails? Awesome!|||Probably the El
9	'I tend to build up a collection of things on my desktop that i use frequently and then move them into a folder called 'Everything' from there it get sorted into type and sub type|||i ike to collect odd objects, even at w
10	I'm not sure, that's a good question. The distinction between the two is so dependant on perception. To quote Robb Flynn, "The hate you feel is nothing more, than love you feel to win this war."|||Good question! It'
11	'https://www.youtube.com/watch?v=w8-egj0y8Qs|||I'm in this position where I have to actually let go of the person, due to a various reasons. Unfortunately I'm having trouble mustering enough strength to actually
12	'One time my parents were fighting over my dad's affair and my dad pushed my mom. The fall broke her finger. She's pointed a gun at him and made him get on his knees and beg for his life. She's...|||I'm gonna ta
13	'https://www.youtube.com/watch?v=PLAaiKvHvZs|||51 :o|||I went through a break up some months ago. We were together for 4 years and I had planned my life around that relationship. I wasn't the one breaking t
14	'Joe santagato - ENTP|||ENFJ or ENTP? I'm not too sure of his type yet|||You know you're not INFJ if heavy Fi doesn't make you want to violently bang your head against a wall lol You know you're not INFJ if you d
15	'Fair enough, if that's how you want to look at it. Like I stated before, they were incredibly naive in their comments... However, they think those are things that would help us because those are the...|||For myself, ME

Figure 3. 3 Dataset without MBTI personality labeling

Consequently, the dataset was uploaded to Google Drive to be processed by using the google Colab. Figure 3.4 represents the dataset view after Google Drive mounting and reading the dataset.

	0	1	2	3	4	5	6
0	'http://www.youtube.com/watch?v=qsXHcwe3krw	http://41.media.tumblr.com/tumblr_lfouy03PMA1q...	enfp and intj moments https://www.youtube.com...	What has been the most life-changing experienc...	http://www.youtube.com/watch?v=vXZeYwwRDw8 h...	May the PerC Experience immerse you.	The last thing my INFJ friend posted on his fa...
1	'I'm finding the lack of me in these posts ver...	Sex can be boring if it's in the same position...	Giving new meaning to 'Game' theory.	Hello *ENTP Grin* That's all it takes. Than w...	This + Lack of Balance and Hand Eye Coordination.	Real IQ test I score 127. Internet IQ tests ar...	You know you're an ENTP when you vanish from a...
2	'Good one _____ https://www.youtube.com/wat...	Of course to which I say I know; that's my ble...	Does being absolutely positive that you and yo...	No I didn't; thank you for a link!	So-called Ti-Si loop (and it can stem from any...	Have you noticed how peculiar vegetation can b...	The Smiths – Never Had No One Ever
3	'Dear INTP I enjoyed our conversation the ot...	Dear ENTJ sub Long time no see. Sincerely A...	None of them. All other types hurt in deep exi...	Probably a sliding scale that depends on indiv...	Draco Malfoy also. I'd say he's either 358 or ...	I'm either 358 or 385 though in which stacking...	I'm not particularly introverted or extraverte...
4	'You're fired.	That's another silly misconception. That appro...	But guys... he REALLY wants to go on a super-d...	Never mind. Just go on permanent vacation.	Two months? I wouldn't be crazy about the idea...	Lol. Its not like our views were unsolicited. ...	Sometimes I just really like impoverished rap ...

Figure 3.4 Dataset view from Google Drive.

Now changing the Data Frame to a hierarchical index structure by converting the columns into rows. After stacking the Data Frame, an operation resets the index levels, turning the index levels into columns and generating a default integer index beginning in 0. This effectively converts the hierarchical index structure back into a regular columnar structure while retaining the stacked data. A process of reshaping the original Dataset from a wide format (with multiple columns) to a long format (with a multi-level index), and then it resets the index levels to turn them into columns, resulting in a flattened Dataset with a more conventional structure.

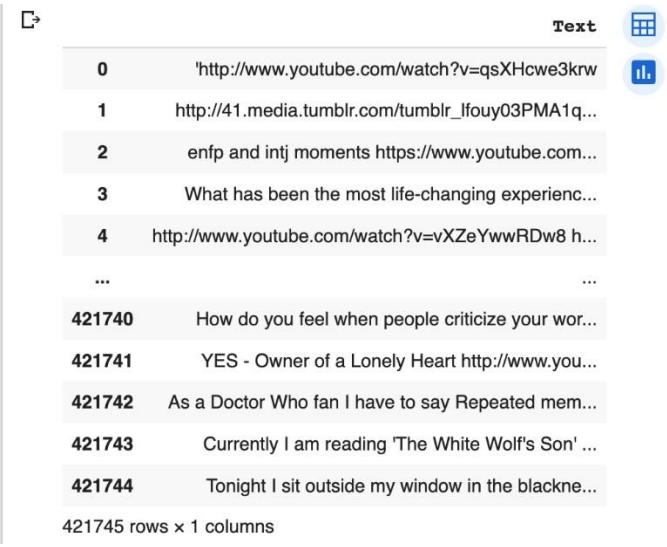
One of the most important dataset-cleaning steps is removing the missing values (NaN). The process of cleaning also in the dataset is dropping the columns, not the rows, if all values of the column are NaN, then the overall column will be deleted. Figure 3.5 represents the result of the process.

	level_0	level_1	0
0	0	0	'http://www.youtube.com/watch?v=qsXHcwe3krw
1	0	1	http://41.media.tumblr.com/tumblr_lfouy03PMA1q...
2	0	2	enfp and intj moments https://www.youtube.com...
3	0	3	What has been the most life-changing experienc...
4	0	4	http://www.youtube.com/watch?v=vXZeYwwRDw8 h...
...	...	...	...
421745	8674	45	I was going to close my facebook a few months ...
421746	8674	46	30 Seconds to Mars - All of my collections. It...
421747	8674	47	I have seen it and i agree. I did actually thi...
421748	8674	48	Ok so i have just watched Underworld 4 (Awaken...
421749	8674	49	I would never want to turn off my emotions. so...

421750 rows × 3 columns

Figure 3.5 Reshaping and indexing the dataset

To simplify the structure of the cleaned dataset, the columns level_0 and level_1 must be removed as these columns were generated by a previous process, See Figure 3.6.



	Text
0	'http://www.youtube.com/watch?v=qsXHcwe3krw
1	http://41.media.tumblr.com/tumblr_lfouy03PMA1q...
2	enfp and intj moments https://www.youtube.com...
3	What has been the most life-changing experienc...
4	http://www.youtube.com/watch?v=vXZeYwwRDw8 h...
...	...
421740	How do you feel when people criticize your wor...
421741	YES - Owner of a Lonely Heart http://www.you...
421742	As a Doctor Who fan I have to say Repeated mem...
421743	Currently I am reading 'The White Wolf's Son' ...
421744	Tonight I sit outside my window in the blackne...

421745 rows × 1 columns

Figure 3.6 One column of the needed tweets.

The next step in the pre-processing stage is the processing of lemmatization of the dataset to reduce the words to their root. This process is for text analysis, and sentiment analysis, to increase the efficiency and accuracy of text processing. Stemming algorithms are used for dataset cleaning, to improve the efficiency of text analysis, search, and retrieval tasks. For the tweeting dataset removing the RT which stands for retweet. Lowering the text to revert all text from any capitalization. Removing the URLs. Remove all special characters except for emojis because they are important in sentiment analysis. Applying the lemmatize and stop words removal.

After all previous processes, a new column containing the cleaned text is created. Also creating an additional column for the tokenized text called sentences. See Figure 3.7.

	Text	clean text	sentences
0	'http://www.youtube.com/watch?v=qsXHcwe3krw		[]
1	http://41.media.tumblr.com/tumblr_lfouy03PMA1q...		[]
2	enfp and intj moments https://www.youtube.com...	enfp intj moment sportscenter top ten play prank	[enfp intj moment sportscenter top ten play pr...
3	What has been the most life-changing experienc...	life changing experience life	[life changing experience life]
4	http://www.youtube.com/watch?v=vXZeYwwRDw8 h...	repeat today	[repeat today]
...	...	...	...
421745	I was going to close my facebook a few months ...	going close facebook month back well wanting a...	[going close facebook month back well wanting ...
421746	30 Seconds to Mars - All of my collections. It...	second mar collection seems fitting mood right	[second mar collection seems fitting mood right]
421747	I have seen it and i agree. I did actually thi...	seen agree actually think first time watched m...	[seen agree actually think first time watched ...
421748	Ok so i have just watched Underworld 4 (Awaken...	ok watched underworld awakening must say reall...	[ok watched underworld awakening must say real...
421749	I would never want to turn off my emotions. so...	would never want turn emotion sometimes hide w...	[would never want turn emotion sometimes hide ...

Figure 3.7 Cleaned and tokenized text.

The next step is labeling the text against the Big Five personality. First initializing the sentiment process. See Figure 3.8 for the code of this process

Labling

```
1  # Initializing Sentiment Analyzer
2  import nltk
3  nltk.download('vader_lexicon')
4  sia = SentimentIntensityAnalyzer()
5  def label_sentiment(sentiment_scores):
6      pos_score = sentiment_scores['pos']
7      neg_score = sentiment_scores['neg']
8      neu_score = sentiment_scores['neu']
9      compound_score = sentiment_scores['compound']
10
11     if compound_score > 0.5 and pos_score > neg_score:
12         return "Conscientiousness"
13     elif neg_score > 0.5:
14         return "Neuroticism"
15     elif pos_score > 0.5:
16         return "Agreeableness"
17     elif compound_score >= 0.5:
18         return "Openness"
19     elif pos_score > neg_score:
20         return "Extraversion"
21     else:
22         return "Neutral"
```

Figure 3.8 Personality labeling.

Lines 4 to 9 of this code are related to sentiment analysis using the VADER (Valence Aware Dictionary and sEntiment Reasoner) sentiment analysis tool, which is available as part of the Natural Language Toolkit (NLTK) library in Python.

Here's what does this code mean:

1. `sia = SentimentIntensityAnalyzer()`: This line creates an instance of the `SentimentIntensityAnalyzer` class from the NLTK library. This class provides a pre-trained model for performing sentiment analysis on text using the VADER sentiment analysis tool.
2. The next part of the code defines a function named `label_sentiment(sentiment_scores)`. This function takes a dictionary of sentiment scores as an argument.
3. Inside the function:

- `pos_score = sentiment_scores ['pos']`: This line extracts the positive sentiment score from the `sentiment_scores` dictionary. The `''pos''` key corresponds to the positive sentiment score.
- `neg_score = sentiment_scores['neg']`: This line extracts the negative sentiment score from the `sentiment_scores` dictionary. The `''neg''` key corresponds to the negative sentiment score.
- `neu_score = sentiment_scores['neu']`: This line extracts the neutral sentiment score from the `sentiment_scores` dictionary. The `''neu''` key corresponds to the neutral sentiment score.
- `compound_score = sentiment_scores['compound']`: This line extracts the compound sentiment score from the `sentiment_scores` dictionary. The `''compound''` key corresponds to the compound sentiment score, which is a single score that reflects the text's overall sentiment intensity. Both the positive and negative feelings are considered.

This code initializes a sentiment analysis tool using VADER (`sia`), defines a function to extract different sentiment scores from a given dictionary, and assigns these scores to individual variables (`pos_score`, `neg_score`, `neu_score`, and `compound_score`) for further analysis or interpretation. VADER's compound score is often used to determine the overall sentiment of a text, as it combines both positive and negative sentiments.

Lines 11 to 22 of this code provide a decision structure that categorizes the sentiment of a given text of the dataset into one of several personality traits based on certain conditions as below. It interprets sentiment scores obtained from a sentiment analysis tool (VADER) and then maps them to various personality traits using a set of rules:

1. If the `compound_score` is greater than 0.5 and the positive sentiment score (`pos_score`) is greater than the negative sentiment score (`neg_score`), the text is labeled as "Conscientiousness".
2. If the `neg_score` is greater than 0.5, the text is labeled as "Neuroticism".
3. If the `pos_score` is greater than 0.5, the text is labeled as "Agreeableness".
4. If the `compound_score` is greater than or equal to 0.5, the text is labeled as "Openness".
5. If the positive sentiment score (`pos_score`) is greater than the negative sentiment score (`neg_score`), the text is labeled as "Extraversion".
6. If none of the above conditions are met, the text is labeled as "Neutral".

Each of the personality traits represents a different aspect of an individual's personality based on their linguistic sentiment in the text.

The next step is to apply the sentiment analyzer to the cleaned text and create a new column called label representing the labeling of the personality trait and saving it to BigPersonalitdata.csv dataset. See Figure 3.9.

```
1 df
```

	Text	clean text	sentences	sentiment	label
0	'http://www.youtube.com/watch?v=qsXHcwe3krw		[]	{'neg': 0.0, 'neu': 0.0, 'pos': 0.0, 'compound...	Neutral
1	http://41.media.tumblr.com/tumblr_lfouy03PMA1q...		[]	{'neg': 0.0, 'neu': 0.0, 'pos': 0.0, 'compound...	Neutral
2	enfp and intj moments https://www.youtube.com...	enfp intj moment sportscenter top ten play prank	[enfp intj moment sportscenter top ten play pr...	{'neg': 0.0, 'neu': 0.588, 'pos': 0.412, 'comp...	Extraversion
3	What has been the most life-changing experienc...	life changing experience life	[life changing experience life]	{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound...	Neutral
4	http://www.youtube.com/watch?v=vXZeYwwRDw8h...	repeat today	[repeat today]	{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound...	Neutral
...	...	...	...	...	...
421745	I was going to close my facebook a few months ...	going close facebook month back well wanting a...	[going close facebook month back well wanting ...	{'neg': 0.0, 'neu': 0.725, 'pos': 0.275, 'comp...	Conscientiousness
421746	30 Seconds to Mars - All of my collections. It...	second mar collection seems fitting mood right	[second mar collection seems fitting mood right]	{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound...	Neutral
421747	I have seen it and i agree. I did actually thi...	seen agree actually think first time watched m...	[seen agree actually think first time watched ...	{'neg': 0.0, 'neu': 0.872, 'pos': 0.128, 'comp...	Extraversion
421748	Ok so i have just watched Underworld 4 (Awaken...	ok watched underworld awakening must say reall...	[ok watched underworld awakening must say real...	{'neg': 0.0, 'neu': 0.62, 'pos': 0.38, 'compou...	Conscientiousness
421749	I would never want to turn off my emotions. so...	would never want turn emotion sometimes hide w...	[would never want turn emotion sometimes hide ...	{'neg': 0.268, 'neu': 0.732, 'pos': 0.0, 'comp...	Neutral

421750 rows × 5 columns

Figure 3.9 Data labeled against Big Five personality traits.

3.5 Splitting Data

Each dataset is split using a stratified approach into a 30% testing dataset (the unseen dataset) and a 70% training dataset. The unseen dataset is used to evaluate the deep learning models after they have been trained on the training dataset. Figure 3.10 represents the process of counting the values in the label of the dataset. Accordingly, Figure 3.11 represents the process of excluding the neutral labelled values, Figure 3.12 represents the test dataset (the unseen dataset), and Figure 3.13 represents the training dataset.

```
Conscientiousness    154417
Neutral              140886
Extraversion          90944
Neuroticism           11770
Agreeableness          9609
Openness                 25
Name: label, dtype: int64
```

Figure 3.10 Values of the labeled column

```
Conscientiousness    154417
Extraversion          90944
Neuroticism           11770
Agreeableness          9609
Openness                 25
Name: label, dtype: int64
```

Figure 3.11 The Big Five personalities and their values

	Unnamed: 0	Text	clean text	sentences	sentiment	label
48107	48107	Are you a fuckin nutjob? Have you ever been to...	fuckin nutjob ever crazy home slice wrist kill...	['fuckin nutjob ever crazy home slice wrist ki...	{'neg': 0.627, 'neu': 0.299, 'pos': 0.075, 'co...	Neuroticism
141135	141135	There are things about INTP's that really driv...	thing intp really drive nut like thinking alou...	['thing intp really drive nut like thinking al...	{'neg': 0.0, 'neu': 0.743, 'pos': 0.257, 'comp...	Conscientiousness
288788	288788	You've every right to be pissed with her. If I...	every right pissed situation confront friend f...	['every right pissed situation confront friend...	{'neg': 0.362, 'neu': 0.245, 'pos': 0.393, 'co...	Extraversion
334288	334288	Atheist yes. Vegan/vegetarian no. Bisexual yes...	atheist yes vegan vegetarian bisexual yes libe...	['atheist yes vegan vegetarian bisexual yes li...	{'neg': 0.0, 'neu': 0.645, 'pos': 0.355, 'comp...	Conscientiousness
219666	219666	This was a perfect thread for me to read right...	perfect thread read right	['perfect thread read right']	{'neg': 0.0, 'neu': 0.448, 'pos': 0.552, 'comp...	Conscientiousness
...	...	...	...	...	...	...
164484	164484	Henry Tilney is my favourite Austen hero too. ...	henry tilney favourite austen hero feel undera...	['henry tilney favourite austen hero feel unde...	{'neg': 0.0, 'neu': 0.393, 'pos': 0.607, 'comp...	Conscientiousness
214439	214439	Dancing.. no thank you.	dancing thank	['dancing thank']	{'neg': 0.0, 'neu': 0.286, 'pos': 0.714, 'comp...	Agreeableness
302001	302001	Gulp. I actually texted the girl I like for th...	gulp actually texted girl like first time idio...	['gulp actually texted girl like first time id...	{'neg': 0.149, 'neu': 0.588, 'pos': 0.262, 'co...	Extraversion
333467	333467	Sorry It's just that you've had an account lon...	sorry account longer thought would know make t...	['sorry account longer thought would know make...	{'neg': 0.072, 'neu': 0.611, 'pos': 0.317, 'co...	Conscientiousness
379594	379594	Ugh. I hate that I identify with this so much...	ugh hate identify much romanticeditor original...	['ugh hate identify much romanticeditor origin...	{'neg': 0.305, 'neu': 0.329, 'pos': 0.366, 'co...	Extraversion

53353 rows × 6 columns

Figure 3.12 Testing dataset

	Unnamed: 0	Text	clean text	sentences	sentiment	label
177135	177135	If someone asks me to go out and I say no if t...	someone asks go say next word free food buying...	['someone asks go say next word free food buyi...	{'neg': 0.0, 'neu': 0.752, 'pos': 0.248, 'comp...	Conscientiousness
81197	81197	I can always count on myself to think that my ...	always count think emotion extremely important...	['always count think emotion extremely importa...	{'neg': 0.0, 'neu': 0.672, 'pos': 0.328, 'comp...	Conscientiousness
151989	151989	Good old INFJs. I can usually count on the INF...	good old infjs usually count infjs know spirit...	['good old infjs usually count infjs know spir...	{'neg': 0.0, 'neu': 0.672, 'pos': 0.328, 'comp...	Conscientiousness
259162	259162	Because my girlfriend looks just like her (onl...	girlfriend look like dark hair	['girlfriend look like dark hair']	{'neg': 0.0, 'neu': 0.615, 'pos': 0.385, 'comp...	Extraversion
342500	342500	yes yes it is.	yes yes	['yes yes']	{'neg': 0.0, 'neu': 0.0, 'pos': 1.0, 'compound...	Conscientiousness
...	...	...	...	...	...	...
256033	256033	She just sounds like she's PMSing. Are you sur...	sound like pmsing sure friendship wink	['sound like pmsing sure friendship wink']	{'neg': 0.0, 'neu': 0.28, 'pos': 0.72, 'compou...	Conscientiousness
87025	87025	I always wear a jacket or cardigan. It's a com...	always wear jacket cardigan compulsive thing l...	['always wear jacket cardigan compulsive thing...	{'neg': 0.0, 'neu': 0.762, 'pos': 0.238, 'comp...	Conscientiousness
338238	338238	Sounds pretty much like a combination of me an...	sound pretty much like combination best friend...	['sound pretty much like combination best frie...	{'neg': 0.0, 'neu': 0.234, 'pos': 0.766, 'comp...	Conscientiousness
194606	194606	If they are cold to you then it doesn't seem l...	cold seem like right person	['cold seem like right person']	{'neg': 0.0, 'neu': 0.615, 'pos': 0.385, 'comp...	Extraversion
96632	96632	My photos are usually pretty normal I think. ...	photo usually pretty normal think	['photo usually pretty normal think']	{'neg': 0.0, 'neu': 0.556, 'pos': 0.444, 'comp...	Extraversion

213412 rows × 6 columns

Figure 3.13 Training dataset

3.6 Deep Learning Modeling

By using the Word Embedding Technique to convert the text data (words) into vectors, every word is represented as an n-dimensional dense vector, with victors that are comparable for similar words. Pretrained Word2Vec is used for word embedding to extract features. Word2Vec is a technique for natural language processing NLP was published in the year 2013. This technique algorithm uses a natural network model to learn word associations from a large frame of text. In the time of training, the model can detect synonymous words or give suggestions of other words in a part of a sentence. As the name represents, Word2Vec represents every distinct feature of a word with a particular list of numbers called vectors. These vectors were chosen carefully to capture the needed semantic and syntactic feature qualities of words. In this research, Word2Vec is used as a group of related models to be able to produce word embedding. These models are two neural networks that have been trained to reconstruct the linguistic context of the group of words. This Word2vec model takes the inputs of a large corpus of texts to produce vector space, it generates several hundred dimensions. So, each unique word in the frame or corpus is assigned to a corresponding vector in the entire text.

Word2Vec is a popular technique in natural language processing (NLP) that aims to capture semantic relationships between words by training on large text corpora. The "negative300" in the filename indicates that the word vectors have been trained with a dimensionality of 300, meaning each word is represented as a vector of 300 numerical values. These pre-trained word vectors have proven to be highly useful in a variety of NLP tasks, such as text classification, sentiment analysis, and machine translation.

In this study, there are two models used. The first model is the LSTM. It is for feature extraction in the context of predicting Big Five personality traits, involves training the model on textual data associated with individuals' personality traits and then using the learned features from the model for personality trait prediction. The LSTM model does word embedding by converting the processed words into numerical vectors called word embedding. As mentioned previously the pre-trained word embedding used is Word2Vec. After reading the training and testing data, the process of converting labels to categorical, as in Figure 3.14, is a common process when dealing with a classification task. The importance of this conversion is to allow the model to understand and process the labels effectively.

1.

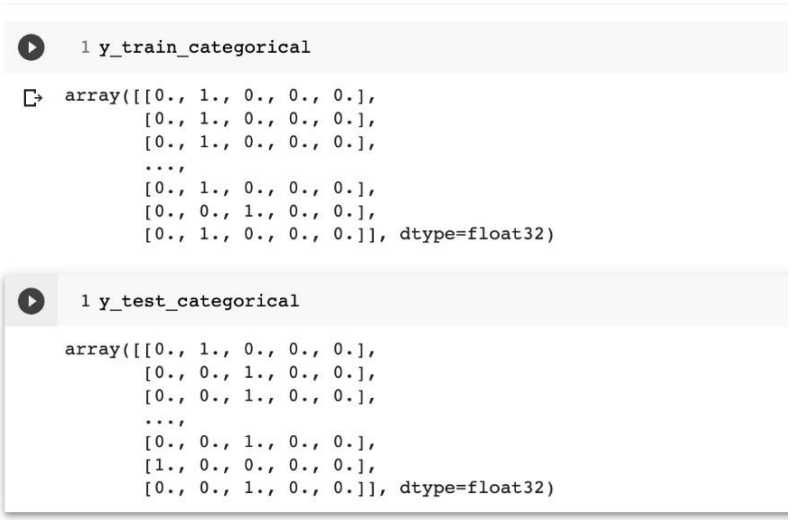


Figure 3.14 Converting label to categorical

The process of converting the x_train and X_test into numbers is essential because machine learning algorithms operate on numerical data. Therefore, text data needs to be transformed into a numerical data format so the model can understand and process effectively. Figure 3.15 represents the output of the converting x_train and X_test into numbers. This process is for preparing the converted text to numerical as an input to the LSTM model.

```
Total unique tokens generated:  64635
Shape of padded train tensor:  (213481, 142)
Shape of padded test tensor:  (53371, 142)
```

Figure 3.15 Converting text to numerical

In this process, the total unique tokens resulted were 64635 words or sub-words across all the text in the dataset. Each unique token represents a distinct word or piece of text in the dataset. Shape of padded train tensor (213481, 142). The dataset is organized in tensors as tensors are multidimensional arrays. The shape (213481, 142) refers to the shape of the training data after it has been preprocessed and padded for use in the LSTM model. 213481 is the number of records in the training sample, where 142 is the length of each sequence after padding. This means that each input sequence (text) has been padded or truncated to have a max length of 142 tokens. In this study, 142 was chosen because the limit of the tweet is 142 characters and this is necessary for inputting them into the LSTM model. The shaping of padded test tensor (53371, 142). The number 53371 is for the testing sample. 142 is the length of each sequence after padding.

The total unique tokens generated provide insight into the vocabulary size of the dataset. The shape of the padded training tensor and the shape of the padded test tensor provide how the dataset has been structured and prepared for the LSTM model. Having consistent sequence length due to the padding is a very important process for efficient batch processing and model training. These steps are part of the dataset processing pipeline. Where a process of

transforming raw text data into a format that the model LSTM can accept and process efficiently. It involves converting words into numerical embeddings, creating sequences of fixed length, and organizing data into tensors that can be used as input for the training and evaluation.

Building an embedding matrix for initializing the embedding layer in the LSTM model to allow the model to leverage the semantic relationships and word meaning captured by this embedding. This is for providing the LSTM model with a foundation of a semantic understanding, which can significantly enhance its ability to extract meaningful information from the input text data.

For the best execution time google Colab GPUs were added to it. The new GPU's cut the speed of the execution time from hours to minutes. So, the execution time for each epoch is 18 seconds which is very fast running. Figure 3.16 shows the LSTM execution time.

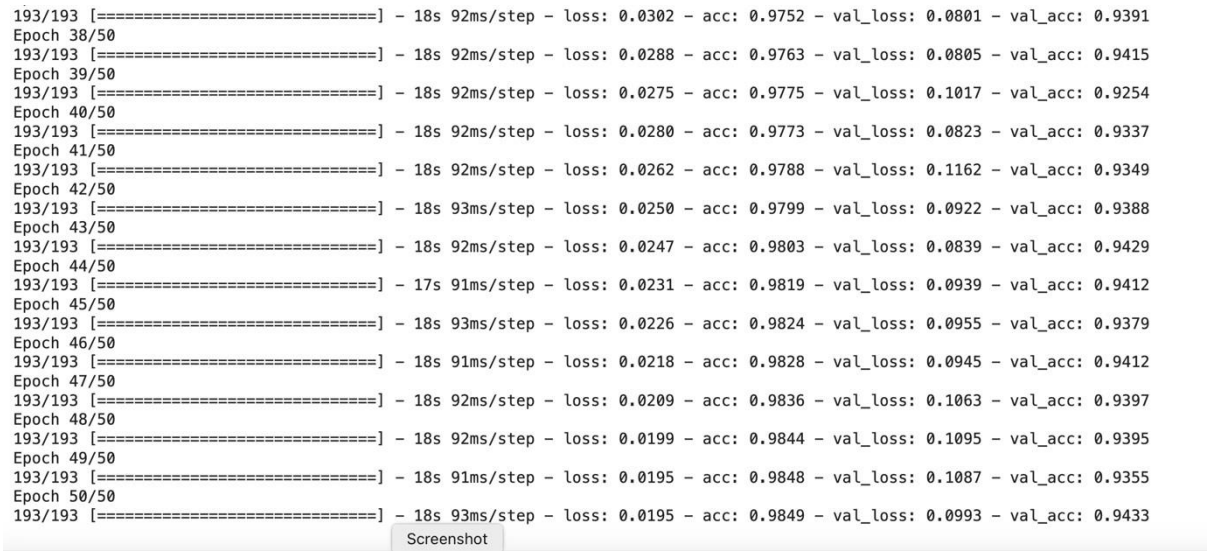


Figure 3.16 LSTM execution time

Optimize and build the LSTM model. The construction of the sequential model using Keres for the text classification task incorporates the LSTM model architecture. The LSTM model sequentially integrates an embedding layer initialized with the pre-trained word embeddings,

the LSTM layer has 850 units for sequence analysis, a dropout layer a dense layer with SoftMax activation to forecast one of the Big Five classes, in order to avoid overfitting. Binary crossentropy loss and accuracy are used as the evaluation metrics for building the model. This configuration of the model is for processing the sequence of text data and predicting the categorical outcomes of the multi-class personality trait classification.

The tuning of a training procedure for an LSTM-based text categorization model. The training procedure is arranged to repeat over the dataset a certain number of times by specifying an epoch count of 50, enabling the model to gradually learn from the data. The number of instances used in each weight adjustment iteration is controlled by the intended batch size, set at 1000, which has an impact on convergence efficiency. It is noteworthy that the 'EarlyStopping' callback mechanism with patience of 30 epochs has been integrated to reduce overfitting. If the validation accuracy does not improve after a certain number of subsequent epochs, this callback stops the training process, preventing the undesired specialization of the model. Together, these modifications optimize the training trajectory of the LSTM model by combining controlled batch processing, systematic epoch management, and adaptive stopping conditions to improve model performance without overfitting.

As LSTM-based text categorization models, the training phase is represented by the code extract that is being provided. The model is trained using the prepared training data, padded_train, and its associated one-hot encoded categorical labels, designated as "y_train_categorical," using the "fit" function from the Keras framework. The 'epoch' option specifies the number of times the training procedure is iterated over a certain number of epochs. The granularity of weight alterations made during optimization is determined by the selected batch size, denoted as 'BATCH_SIZE'. In addition, a validation split of 10% is used to evaluate how well the model generalizes during the training cycle. To aid in the examination of model performance, metrics such as accuracy, training and validation loss, and training loss are

tracked. The resulting "history" object contains these metrics for later assessment. In conclusion, this code orchestration works to train the LSTM model using the supplied training data, giving it the capacity to recognize underlying patterns and relationships within textual data for precise categorization.

The LSTM model has been trained using the Keras library. Use the save method to store the model's architecture, learned weights, and configuration after importing the required function. The model is stored as a file in the HDF5 (.h5) format. This method guarantees that the trained LSTM model can be used again without having to be retrained. The model's fundamental characteristics are contained in the saved model file, making it easy to reload the model for tasks requiring prediction, adjustment, or evaluation. By saving the model, the learned architecture, configuration, and weights are stored. The model can then be loaded and used for evaluation, fine-tuning, or forecasts without having to start from scratch again.

The second model is the Gated Recurrent Unit (GRU) model. This model is dedicated to capturing patterns and relationships within sequential data. The key feature of the GRU model is its gating mechanism, which allows it to selectively update and store information in a hidden state, GRU is designed to be more computationally efficient and have fewer parameters than traditional LSTM models. GRU model is particularly useful when dealing with sequential data, such as text or time series. The model can capture long-range dependencies in sequences while avoiding some of the complexities of the traditional LSTM model. In addition, this model has gained popularity in NLP tasks like machine translation, sentiment analysis, and language generation, where understanding context and relationships between words over varying time scales is very important. The model has the same procedures and processes as the LSTM model, reading training dataset, reading test dataset, converting the label to categorical, converting x_train and X_test into numbers, embedding model, and defining functions that are used as a matrix. There are three custom evaluation metrics in this mode, the `recall_m`, `precision_m`,

and `f1_m`. They are tailored for binary classification tasks within TensorFlow-based models. The `recall_m` function computes the recall or true positive rate, assessing the model's ability to correctly identify positive instances, the `precision_m` function calculates precision, measuring the accuracy of positive predictions, and the `f1_m` function determines the F1score, a harmonic mean of precision and recall, offering a balanced performance measure. These custom metrics are designed to address imbalanced class scenarios and can be integrated into the model evaluation process during training and validation in Keras, providing insights into the model's performance with respect to true positive predictions, accurate positive predictions, and their harmonized combination. The last step is optimizing and building the model. The running time for GRU epochs are 17 and 18 seconds as the paid Colab GPUs. Figure 3.17 shows the execution time of the GRU model

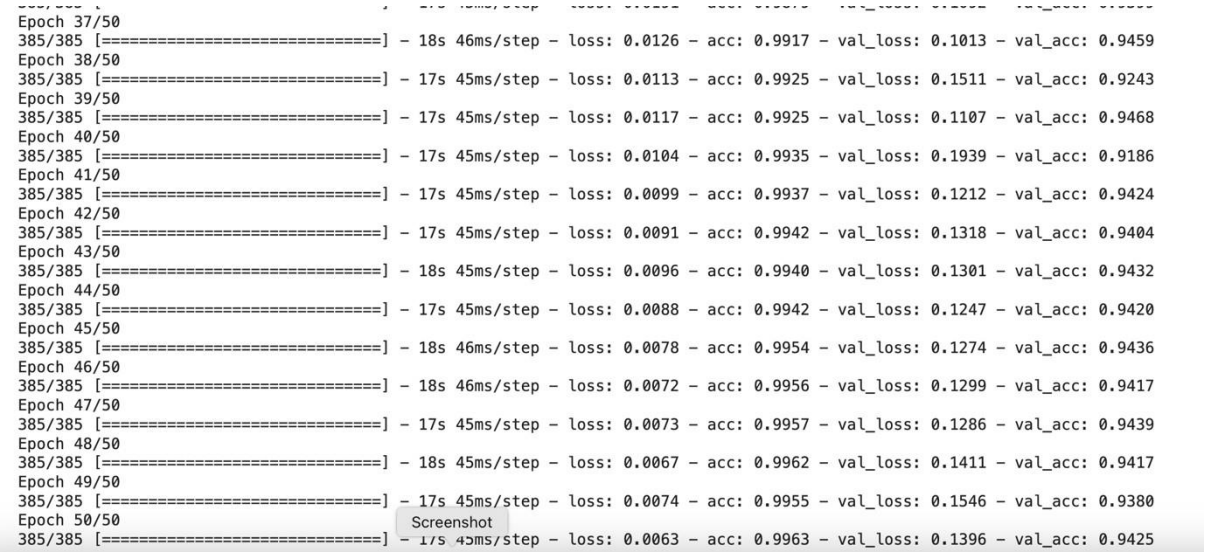


Figure 3.17 GRU model execution time

The RNN model were added to the research. Figure 3.18 shows the execution time for it. The running time per epoch for the RNN Colab between 339 to 354 seconds where free GPUs used.

```
Epoch 1/10
97/97 [==============================] - 354s 4s/step - loss: 0.7420 - acc: 0.4005 - val_loss: 0.3429 - val_acc: 0.5807
Epoch 2/10
97/97 [==============================] - 330s 3s/step - loss: 0.4066 - acc: 0.5248 - val_loss: 0.3400 - val_acc: 0.5809
Epoch 3/10
97/97 [==============================] - 328s 3s/step - loss: 0.5036 - acc: 0.4759 - val_loss: 0.3391 - val_acc: 0.5807
Epoch 4/10
97/97 [==============================] - 326s 3s/step - loss: 0.3558 - acc: 0.5648 - val_loss: 0.3276 - val_acc: 0.5825
Epoch 5/10
97/97 [==============================] - 352s 4s/step - loss: 0.3379 - acc: 0.5796 - val_loss: 0.3263 - val_acc: 0.5799
Epoch 6/10
97/97 [==============================] - 339s 3s/step - loss: 0.3257 - acc: 0.5913 - val_loss: 0.3111 - val_acc: 0.5793
Epoch 7/10
97/97 [==============================] - 340s 4s/step - loss: 0.4599 - acc: 0.5223 - val_loss: 0.3604 - val_acc: 0.5770
Epoch 8/10
97/97 [==============================] - 348s 4s/step - loss: 0.3941 - acc: 0.5356 - val_loss: 0.2830 - val_acc: 0.6108
Epoch 9/10
97/97 [==============================] - 352s 4s/step - loss: 0.2716 - acc: 0.6479 - val_loss: 0.2445 - val_acc: 0.6823
Epoch 10/10
97/97 [==============================] - 354s 4s/step - loss: 0.2319 - acc: 0.7231 - val_loss: 0.2141 - val_acc: 0.7494
```

Figure 3.18 RNN execution time

In the RNN execution run time we can see that it was running on the default Colab run time. Because the GPU's program was ended.

Chapter Four: Results and Discussion

4.1 Results and Evaluation of the Model

Four standard metrics were used to evaluate the DL models, i.e., accuracy, precision, recall, and F1-score, where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives (see Equations 1–4):

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (1)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (2)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (3)$$

$$\text{F1} - \text{Score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

Table 4,4 provides insights about each personality of the Big Five personalities of the tweets in the dataset. The table shows 4 types of personalities, Conscientiousness, Extraversion, Neuroticism, and Agreeableness. The table does not show the Openness personality because of the small number of words that lead to Openness personality in the dataset.

Table (4,4). Insights of each Big Five personality

	LSTM				**GRU**		
Personality types	**Precision**	**Recall**	**F1-score**		**Precision**	**Recall**	**F1-score**
Conscientiousness	92	84	92		90	84	87
Extraversion	94	97	96		96	96	96
Neuroticism	93	90	91		91	93	92
Agreeableness	99	95	97		99	93	96

	RNN		
Personality types	**Precision**	**Recall**	**F1-score**
Conscientiousness	0.73	0.63	0.68
Extraversion	0.81	0.80	0.80
Neuroticism	0.62	0.66	0.64
Agreeableness	0.82	0.60	0.70

4.2 Discussion and Results

From Table 4,4. comparing precision, recall, and F1-scores for LSTM and GRU models in predicting different personality types, several insights can be drawn:

1. Consistency in Performance: Both LSTM and GRU models demonstrate consistent performance across most personality types, as indicated by similar precision, recall, and F1-scores. This suggests that both architectures are effective in capturing patterns related to personality traits.

2. Extraversion Superiority: For the "Extraversion" personality type, both LSTM and GRU models exhibit high precision, recall, and F1-scores, with F1-scores of 96 for both architectures. This indicates that the models excel in identifying instances associated with this trait, suggesting that extraversion-related language patterns are distinct and recognizable.

3. Distinct Characteristics: The "Agreeableness" personality type showcases high precision for both architectures, indicating that when the models predict agreeableness, they are confident in their predictions. However, the recall is comparatively lower for the GRU model (93) than for the LSTM model (95). This suggests that while the GRU model might be more selective in predicting agreeableness, the LSTM model captures more instances associated with this trait.

4. Balanced Performance: Overall, the F1-scores are generally high across the personality types for both LSTM and GRU models. This indicates a balanced trade-off between precision and recall, which is very important in achieving a harmonious blend of accurate positive predictions and thorough coverage of actual positive instances.

5. Differential Performance: The "Neuroticism" personality type showcases a slight difference in performance between LSTM and GRU models. The LSTM model has a

higher precision (93) compared to the GRU model (91), while the GRU model has a slightly higher recall (93) compared to the LSTM model (90). This suggests that the LSTM model might be more cautious in predicting neuroticism, focusing on instances it is confident about, whereas the GRU model captures a broader range of instances related to this trait.

6. Conservative vs. Inclusive: In the case of "Conscientiousness," the LSTM model has slightly higher precision (92) compared to the GRU model (90), indicating that the LSTM model is more conservative in its positive predictions. However, both models show similar recall (84) and F1-scores (92 and 87 for LSTM and GRU, respectively).

7. Model Comparability: Across most personality types, the precision, recall, and F1scores are in close alignment for both LSTM and GRU models. This suggests that in

   this specific analysis, the two architectures perform relatively similarly in terms of personality trait prediction.

8. And as the results in the RNN model shows. It is not that a strong in extracting the personality features as in the LSTM and GRU models.

In summary, both LSTM and GRU models demonstrate competitive performance in predicting personality types. While there are slight differences in precision, recall, and F1-scores, these differences are nuanced and might relate to the models' inherent characteristics in capturing different patterns within the data. The high F1-scores across most traits indicate that the models achieve a balanced compromise between precision and recall, leading to accurate and comprehensive predictions.

Table 4.5 provides insights into the accuracy of each used model, the LSTM and GRU. With the Precision, Recall, and F1-Score for each model.

Table (4.5) Insights for models' accuracy

Model	Accuracy	Precision	Recall	F1-Score
LSTM	93.91	93.89	93.91	93.87
GRU	94.07	94.09	94.07	94.07
RNN	73.7	74.1	73.7	73.79

From Table 4.5 comparing accuracy, precision, recall, and F1-scores for LSTM and GRU models, several insights can be gleaned:

1. Comparable Performance: Both LSTM and GRU models exhibit strong performance, with accuracy, precision, recall, and F1-scores close to or above 93%. This suggests that both architectures are effective in capturing patterns and relationships within the data, leading to consistent and reliable predictions.

2. Marginally Higher Performance of GRU: The GRU model has slightly higher accuracy (94.07%) compared to the LSTM model (93.91%), indicating that the GRU model made correct predictions on a slightly larger portion of the dataset. This could signify that the GRU architecture excels in capturing complex dependencies in the data.

3. Balanced Precision and Recall: Both models show nearly identical precision and recall values, with recall being slightly higher than precision. This balance indicates that the models are making accurate positive predictions while also effectively capturing a substantial portion of the true positive instances.

4. High F1-Scores: The F1-scores for both models are also high (above 93%). The F1score, being the harmonic mean of precision and recall, indicates a harmonious blend of accurate positive predictions and comprehensive identification of actual positive instances. This suggests that both models are capable of achieving a good balance between precision and recall.

5. Consistency in Performance Metrics: The minimal discrepancies between accuracy, precision, recall, and F1-scores for LSTM and GRU models suggest that the models achieve a cohesive and uniform performance across multiple evaluation metrics. This uniformity reinforces the reliability of the model's predictions.

6. Overall Robustness: The similarity in performance between the LSTM and GRU models indicates that both architectures are suitable choices for the specific task. This could be particularly advantageous, as it provides the opportunity to choose between the architectures based on factors such as computational efficiency or ease of training.

7. Fine-Tuning Potential: Given the overall high performance of both models, there might be potential for further enhancing their performance through hyperparameter tuning, feature engineering, or data augmentation techniques. This could lead to even more accurate and robust models.

8. As tables 4.4 and 4.5 the results for the LSTM and the GRU are in 90s but for the RNN model the results are in 70s. So, the RNN model wasn’t strong in the processing as the first two models.

In summary, the LSTM and GRU models demonstrate highly comparable and competitive performance in terms of accuracy, precision, recall, and F1-scores. The minor performance differences suggest that both architectures effectively capture the underlying patterns in the

data. The results highlight the potential of both models for accurate prediction tasks and lay the groundwork for potential optimization and further analysis.

Figure 4.19 provides insights into the average of the performance for each model and the evaluation matrix. The GRU model provides a better average than the LSTM.

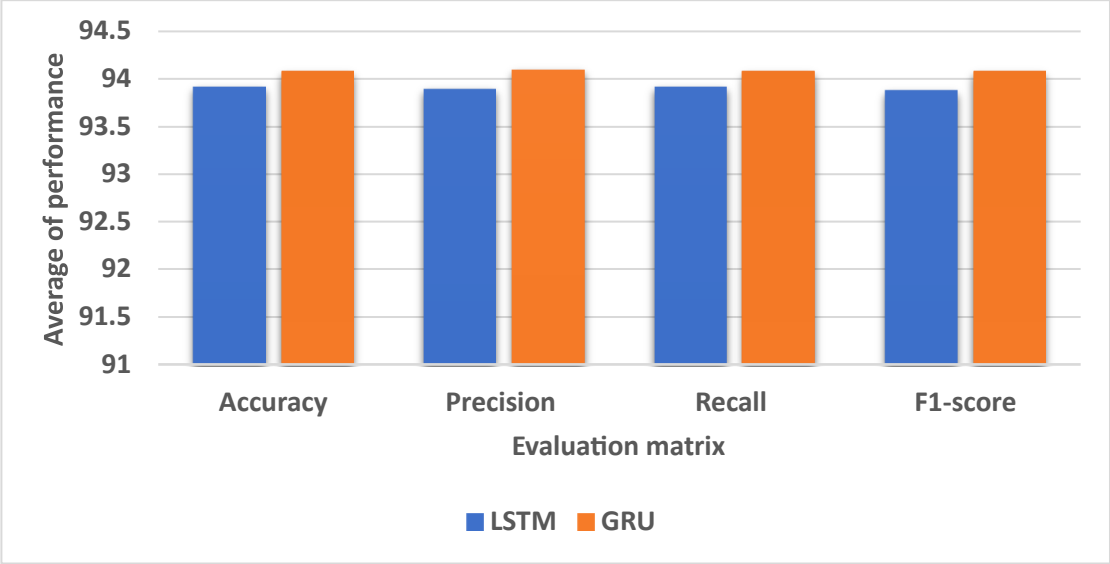


Figure 4.19 Models average

Figure 4.20 provides the extended ROC Curve to the study classes. The true positive rate with the False Positive rate. As the paragraph, the personalities' Conscientiousness, Extraversion, Neuroticism, and Agreeableness curves are very high. They are between 98% and 100 %. But for the openness the result is very low because no enough words in the dataset.

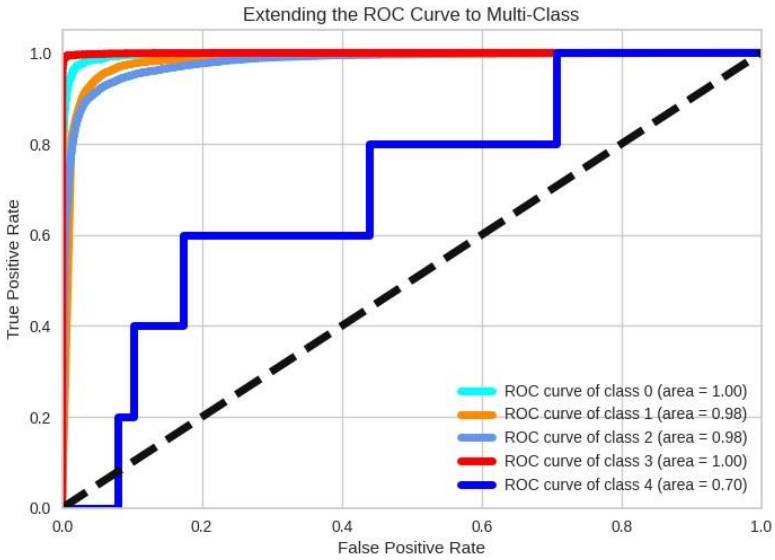


Figure 4.20 The ROC curve

Figure 4.21 provides insight into the frequency of each personality trait indicates.

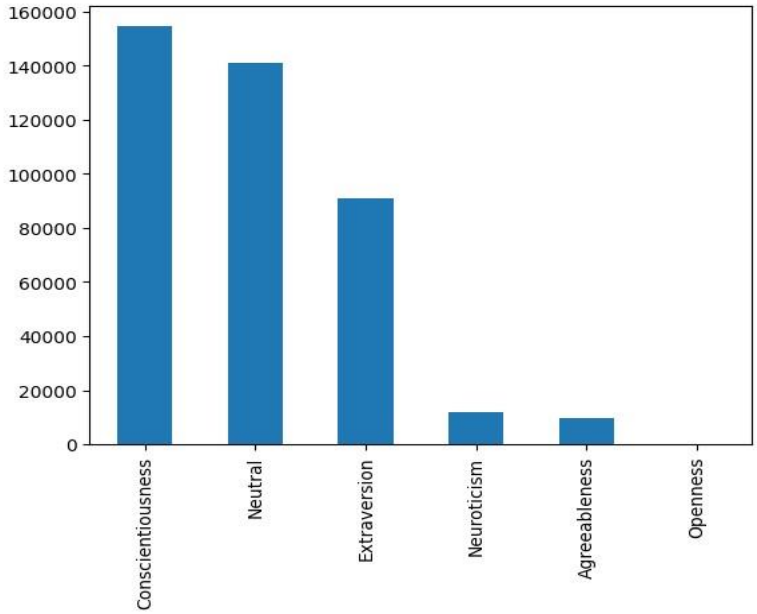


Figure 4.21 Frequency indications

The word cloud figures provide insights into how individuals express themselves and the type of content they share on social media. The dominant words, larger and bolder words represent the traits of the personality. The green color provides the positive sentiments and the brown color represents the negative sentiments. Themes and interests, the words that appear larger reveal the theme and interests associated with each personality trait.

The input is a corpus and documented term matrix. The EDA is the summarizing of the main characteristics of the dataset often using visual methods. The outputs figure out the main trends in the data and if it makes sense. Some ways can be used to explore the dataset as top words to find the most common words for each comedian, vocabulary as looking at the unique number of words used, and the amount of profanity by noting the number of swear words said.

For top words need to matrix, aggregate the data to find the top words, visualize by finding the top 30 words for each comedian, and the insights by having a bunch of word clouds.

Figure 4.22 provides insights into the most repeated words for Conscientiousness personality

Figure 4.22 Conscientious personality

Figure 4.23 provides insights into the Extroversion personality traits

Figure 4.23 Extroversion personality

Figure 4.24 provides insights into the most repeated words for Neuroticism personality

Figure 4.24 Neuroticism personality

Figure 4.25 provides insights into the most repeated words of the Agreeableness personality

Figure 4.25 Agreeableness personality

Figure 4.26 provides insights into the most repeated words for the Openness personality

Figure 4.26 Openness personality

Chapter Five: Conclusions and Future Work

5.1 Overview

In conclusion, the analysis of users' tweets related to the Big Five personality traits offers valuable insights into how individuals utilize social media platforms, notably Twitter, as a means of self-expression and self-reflection.

5.2 Conclusion

Dominant traits, such as Openness and Extraversion, emerge from the word cloud analysis, highlighting the significance of these characteristics in users' self-perception and discussions. Moreover, the presence of emotionally charged terms in the Neuroticism word cloud underscores the platform's role as an outlet for users to share their emotional states and seek support. The identification of specific themes and interests associated with each personality trait suggests the existence of interest-based communities on Twitter, allowing users to engage in discussions aligned with their traits. Ethical and social values associated with Agreeableness are evident, emphasizing the importance of harmonious relationships and ethical behavior in online interactions. These findings lay the groundwork for future research endeavors, including sentiment analysis, topic modeling, user clustering, and longitudinal studies, to gain a deeper understanding of the evolving dynamics of personality expression in the digital age.

5.3 Future Work

There are various exciting areas for future projects and studies such as time series Big Five personality traits clustering and gaining insights as users keep writing and using social media to express themselves. Times series analysis provides immediate personality traits that change as a person's personality changing time to time. With different languages personality mining time series.

References

Al Maruf, H., Meshkat, N., Ali, M. E., & Mahmud, J. (2015a). Human behavior in different social media: A case study of X Platform and Disqus. *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2015*, 270–273. https://doi.org/10.1145/2808797.2809395

Al Maruf, H., Meshkat, N., Ali, M. E., & Mahmud, J. (2015b). Human behavior in different social media: A case study of X Platform and Disqus. *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2015*, 270–273. https://doi.org/10.1145/2808797.2809395

Albayati, M. B., & Altamimi, A. M. (2019a). Identifying fake Facebook profiles using data mining techniques. *Journal of ICT Research and Applications*, *13*(2), 107–117. https://doi.org/10.5614/itbj.ict.res.appl.2019.13.2.2

Albayati, M. B., & Altamimi, A. M. (2019b). Identifying fake Facebook profiles using data mining techniques. *Journal of ICT Research and Applications*, *13*(2), 107–117. https://doi.org/10.5614/itbj.ict.res.appl.2019.13.2.2

Bachrach, Y., Kosinski, M., Graepel, T., Kohli, P., & Stillwell, D. (2012). *Personality and Patterns of Facebook Usage*. www.checkfacebook.com.

Bai, S., Zhu, T., & Cheng, L. (2012a). *Big-Five Personality Prediction Based on User Behaviors at Social Network Sites*. http://arxiv.org/abs/1204.4809

Bai, S., Zhu, T., & Cheng, L. (2012b). *Big-Five Personality Prediction Based on User Behaviors at Social Network Sites*. http://arxiv.org/abs/1204.4809

Bokishev, M. (2022). *Predicting Myers-Briggs Personality Types by the Natural Language from Social Media Posts*. https://www.researchgate.net/publication/361661867

Boulianne, S. (2015). Social media use and participation: a meta-analysis of current research. *Information Communication and Society*, *18*(5), 524–538. https://doi.org/10.1080/1369118X.2015.1008542

Boyd, D. M., & Ellison, N. B. (2007). Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, *13*(1), 210–230. https://doi.org/10.1111/j.1083-6101.2007.00393.x

Carducci, G., Rizzo, G., Monti, D., Palumbo, E., & Morisio, M. (2018). TwitPersonality: Computing personality traits from tweets using word embeddings and supervised learning. *Information (Switzerland)*, *9*(5). https://doi.org/10.3390/info9050127

Celli, F., & Lepri, B. (2019a). Is Big Five better than MBTI? In *Proceedings of the Fifth Italian Conference on Computational Linguistics CLiC-it 2018* (pp. 93–98). Accademia University Press. https://doi.org/10.4000/books.aaccademia.3147

Celli, F., & Lepri, B. (2019b). Is Big Five better than MBTI? In *Proceedings of the Fifth Italian Conference on Computational Linguistics CLiC-it 2018* (pp. 93–98). Accademia University Press. https://doi.org/10.4000/books.aaccademia.3147

Chen, J., Wang, Z., & Liang, Q. (2019). Bandwidth Allocation Based on Personality Traits on Smartphone Usage and Channel Condition. *IEEE Access*, *7*, 102832–102842. https://doi.org/10.1109/ACCESS.2019.2910270

De Montjoye, Y. A., Quoidbach, J., Robic, F., & Pentland, A. (2013a). Predicting personality using novel mobile phone-based metrics. *Lecture Notes in Computer Science (Including*

Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), *7812 LNCS*, 48–55. https://doi.org/10.1007/978-3-642-37210-0_6

De Montjoye, Y. A., Quoidbach, J., Robic, F., & Pentland, A. (2013b). Predicting personality using novel mobile phone-based metrics. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *7812 LNCS*, 48–55. https://doi.org/10.1007/978-3-642-37210-0_6

Dhall, A., & Hoey, J. (2016). First Impressions - Predicting User Personality from X PlatformProfile Images. In M. Chetouani, J. Cohn, & A. A. Salah (Eds.), *Human Behavior Understanding* (pp. 148–158). Springer International Publishing.

Dwivedi, Y. K., Ismagilova, E., Hughes, D. L., Carlson, J., Filieri, R., Jacobson, J., Jain, V., Karjaluoto, H., Kefi, H., Krishen, A. S., Kumar, V., Rahman, M. M., Raman, R., Rauschnabel, P. A., Rowley, J., Salo, J., Tran, G. A., & Wang, Y. (2021). Setting the future of digital and social media marketing research: Perspectives and research propositions. *International Journal of Information Management*, *59*. https://doi.org/10.1016/j.ijinfomgt.2020.102168

EVRİM, V., & NİSSOUL, Y. (2019a). The Relationship among Personality, Interest, and Life Satisfaction of Facebook Users. *Anemon Muş Alparslan Üniversitesi Sosyal Bilimler Dergisi*, *7*, 87–94. https://doi.org/10.18506/anemon.615115

EVRİM, V., & NİSSOUL, Y. (2019b). The Relationship among Personality, Interest, and Life Satisfaction of Facebook Users. *Anemon Muş Alparslan Üniversitesi Sosyal Bilimler Dergisi*, *7*, 87–94. https://doi.org/10.18506/anemon.615115

Faliagka, E., Tsakalidis, A., & Tzimas, G. (2012a). An integrated e-recruitment system for automated personality mining and applicant ranking. *Internet Research*, *22*(5), 551–568. https://doi.org/10.1108/10662241211271545

Faliagka, E., Tsakalidis, A., & Tzimas, G. (2012b). An integrated e-recruitment system for automated personality mining and applicant ranking. *Internet Research*, *22*(5), 551–568. https://doi.org/10.1108/10662241211271545

Farnadi, G., Zoghbi, S., Moens, M. F., & De Cock, M. (2013). Recognizing personality traits using Facebook status updates. *AAAI Workshop - Technical Report*, *WS-13-01*, 14–18. https://doi.org/10.1609/icwsm.v7i2.14470

Felt, M. (2016). Social media and the social sciences: How researchers employ Big Data analytics. *Big Data and Society*, *3*(1). https://doi.org/10.1177/2053951716645828

Gallo, F. R., Simari, G. I., Martinez, M. V., & Falappa, M. A. (2020). Predicting user reactions to X Platform feed content based on personality type and social cues. *Future Generation Computer Systems*, *110*, 918–930. https://doi.org/10.1016/j.future.2019.10.044

Gou, L., Mahmud, J., Haber, E. M., & Zhou, M. X. (2013a). PersonalityViz: A visualization tool to analyze people's personality with social media. *International Conference on Intelligent User Interfaces, Proceedings IUI*, 45–46. https://doi.org/10.1145/2451176.2451191

Gou, L., Mahmud, J., Haber, E. M., & Zhou, M. X. (2013b). PersonalityViz: A visualization tool to analyze people's personality with social media. *International Conference on Intelligent User Interfaces, Proceedings IUI*, 45–46. https://doi.org/10.1145/2451176.2451191

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(15), 5802–5805. https://doi.org/10.1073/pnas.1218772110

Lima, A. C. E. S., & De Castro, L. N. (2013). Multi-label semi-supervised classification applied to personality prediction in tweets. *Proceedings - 1st BRICS Countries Congress on Computational Intelligence, BRICS-CCI 2013*, 195–203. https://doi.org/10.1109/BRICS-CCI-CBIC.2013.41

Liu, D., & Campbell, W. K. (2018). *The Big Five Personality Traits, Big Two Metatraits, and Social Media: A Meta-Analysis*.

Maheshkar, V., & Sarin, S. K. (2022). *Review and Analysis of Emotion Detection from Tweets using X Platform Datasets*.

Mairesse, F., Uk, M. A., Mehl, M. R., & Moore, R. K. (2007). Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text. In *Journal of Artificial Intelligence Research* (Vol. 30).

Markovikj, D., Gievska, S., Kosinski, M., & Stillwell, D. (2013a). *Mining Facebook Data for Predictive Personality Modeling*. http://mypersonality.org/wiki;

Markovikj, D., Gievska, S., Kosinski, M., & Stillwell, D. (2013b). *Mining Facebook Data for Predictive Personality Modeling*. http://mypersonality.org/wiki;

Najm Al-Zaytoonah, N. (2019). Big Five Traits: A Critical Review. In *Article in Gadjah Mada International Journal of Business*. https://www.researchgate.net/publication/335834113 Nguyen, T., Venkatesh, S., Phung, D., & Adams, B. (2011). *Towards Discovery of Influence and*

Personality Traits through Social Link Prediction. Bayesian Optimization View project Representation Learning in Complex Data via Pattern Discovery View project Towards Discovery of Influence and Personality Traits through Social Link Prediction. www.aaai.org

Oberlander, J., & Nowson, S. (2006). *Whose Thumb Is It Anyway? Classifying Author Personality from Weblog Text. Whose thumb is it anyway? Classifying author personality from weblog text*. https://www.researchgate.net/publication/220873652

Ortigosa, A., Carro, R. M., & Quiroga, J. I. (2014). Predicting user personality by mining social interactions on Facebook. *Journal of Computer and System Sciences*, *80*(1), 57–71. https://doi.org/10.1016/j.jcss.2013.03.008

Quercia, D., Kosinski, M., Stillwell, D., & Crowcroft, J. (2011a). Our X Platform profiles, our selves: Predicting personality with Twitter. *Proceedings - 2011 IEEE International Conference on Privacy, Security, Risk and Trust and IEEE International Conference on Social Computing, PASSAT/SocialCom 2011*, 180–185. https://doi.org/10.1109/PASSAT/SocialCom.2011.26

Quercia, D., Kosinski, M., Stillwell, D., & Crowcroft, J. (2011b). Our X Platform profiles, our selves: Predicting personality with Twitter. *Proceedings - 2011 IEEE International Conference on Privacy, Security, Risk and Trust and IEEE International Conference on Social Computing, PASSAT/SocialCom 2011*, 180–185. https://doi.org/10.1109/PASSAT/SocialCom.2011.26

Quercia, D., Lambiotte, R., Stillwell, D., Kosinski, M., & Crowcroft, J. (2012a). The personality of popular Facebook users. *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*, 955–964.

https://doi.org/10.1145/2145204.2145346

Quercia, D., Lambiotte, R., Stillwell, D., Kosinski, M., & Crowcroft, J. (2012b). The personality of popular Facebook users. *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*, 955–964. https://doi.org/10.1145/2145204.2145346

Rao, M., Jayant Kanchugar, P., N, P. M., & Semester, E. (2008). Personality Recognition using Social Media Data. In *International Research Journal of Engineering and Technology*. www.irjet.net

Rijitha R. (2023). *THE IMPACT OF SOCIAL MEDIA MARKETING ON CONSUMER PURCHASE INTENTION*. https://www.researchgate.net/publication/358730635

Samih, A., Fennan, A., & Ghadi, A. (2022). IWVTSA: IMPROVED WORDS VECTORS FOR X PLATFORMSENTIMENTS ANALYSIS The 1st International Conference on Networking, Information Systems & Security View project XEW Big Data Analysis View project IWVTSA: IMPROVED WORDS VECTORS FOR X PLATFORMSENTIMENTS ANALYSIS. *Article in Journal of Theoretical and Applied Information Technology*, *15*, 17. https://www.researchgate.net/publication/365519148

Sayeb, Y., Jebri, M., & Ghezala, H. Ben. (2021). A graph-based recommender system for managing the COVID-19 crisis. *Procedia Computer Science*, *196*, 348–355. https://doi.org/10.1016/j.procs.2021.12.023

Schetgen, L., Bogaert, M., & Van den Poel, D. (2021). *Predicting Donation Behavior: Acquisition Modeling in the Nonprofit Sector Using Facebook Data*.

Shipankar, S., Sawale, G., Shelke, R., & Khairkar, A. (2022). Personality Prediction Using Social Media Platforms. *International Research Journal of Engineering and Technology*. www.irjet.net

Skues, J. L., Williams, B., & Wise, L. (2012). The effects of personality traits, self-esteem, loneliness, and narcissism on Facebook use among university students. *Computers in Human Behavior*, *28*(6), 2414–2419. https://doi.org/10.1016/j.chb.2012.07.012

Stillwell, D., Moens, M.-F., Leuven, K. U., Davalos, S., Farnadi, G., Sitaraman, G., Rohani, M., Kosinski, M., & De Cock, M. (2014a). *How are you doing? Emotions and Personality in Facebook Deep Learning Models for Continual Extraction of Knowledge from Text View project Text-to-image translation using generative adversarial networks View project Golnoosh Farnadi Polytechnique Montréal How are you doing? Emotions and Personality in Facebook*. https://www.researchgate.net/publication/283270667

Stillwell, D., Moens, M.-F., Leuven, K. U., Davalos, S., Farnadi, G., Sitaraman, G., Rohani, M., Kosinski, M., & De Cock, M. (2014b). *How are you doing? Emotions and Personality in Facebook Deep Learning Models for Continual Extraction of Knowledge from Text View project Text-to-image translation using generative adversarial networks View project Golnoosh Farnadi Polytechnique Montréal How are you doing? Emotions and Personality in Facebook*. https://www.researchgate.net/publication/283270667

Tan, D. S., Begole, Bo., Kellogg, Wendy., & SIGCHI (Group: U.S.). (2011). *Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems: 2011 proceeding, Vancouver, BC, Canada - May 07-12, 2011*. ACM Press.

Tarka, P., Kukar-Kinney, M., & Harnish, R. J. (2022). Consumers' personality and compulsive buying behavior: The role of hedonistic shopping experiences and gender in mediatingmoderating relationships. *Journal of Retailing and Consumer Services*, *64*. https://doi.org/10.1016/j.jretconser.2021.102802

Watt, J., Tuke, J., & Mitchell, L. (2023). *Personality Profiling: How informative are social media profiles in predicting personal information?* http://arxiv.org/abs/2309.13065

Simon Kimp, 2021 February, Digital 2021:Palestine, https://datareportal.com/reports/digital2021-palestine.

Simon Kimp, 2022 July, Global Social Media Statistics https://datareportal.com/social-mediausers.

Simon Kimp, 2023 July, Global Social Media Statistics https://datareportal.com/social-mediausers.

Stephanie Heitman, 2022 May, What happens in an internet minute in 2022: 90 fascinating online stats. https://localiq.com/blog/what-happens-in-an-internet-minute/

Thomas International, 2021 March, What are the Big Five personality traits, https://www.thomas.co/resources/type/hr-guides/what-are-big-5-personality-traits.

Stacy Jo Dixon, 2023 October, Global social networks ranked by number of users 2023, https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-ofusers/

Kelland, M. (2017, July 07). Personality Theory. OER Commons. Retrieved October 28, 2023, from https://oercommons.org/authoring/22859-personality-theory.https://www.kaggle.com/datasets/datasnaek/mbti-type?resource=download

الملخص

تؤثر منصات التواصل الاجتماعي بشدة على حياة الناس اليومية. وقد أتاح ذلك لعلماء البيانات فرصة عظيمة لدراستها مع مرور الوقت، بحيث يمكن استخدامها لبيع السلع والأشياء بإعلانات أقوى، مع القدرة على معرفة الأشخاص وتصنيفهم وفقًا لآرائهم السياسية. ويمكن لعلماء البيانات الآن مراقبة تصرفات الأشخاص عن كثب على وسائل التواصل الاجتماعي من خلال تصنيف شخصياتهم أو المعروفة سابقًا باسم تويتر) منصة شائعة حول العالم، حيث) X تجميعها أو تصنيفها. وتُعد منصة يستخدمها ملايين الأشخاص يوميًا. لذا، فهي مكان ضخم للباحثين. ويستخدم هذا البحث تقنيات علم من خلال دراسة الحالة العاطفية لمستخدمي منصة X البيانات لتحليل منشورات المستخدمين على منصة ، وجمع الأشياء المشتركة بين منشوراتهم لفهم العلاقات بين منشوراتهم، والقدرة على تصنيف X شخصياتهم. ومع استخدام التكنولوجيا المتقدمة المتاحة بسهولة في السوق، أصبحت وسائل التواصل الاجتماعي موردًا مهمًا للغاية في مجالات علم البيانات وتحليل البيانات. ومن خلال تويتر، يمكن اكتساب رؤى حول تعبيرات الناس ومشاعرهم وعواطفهم من خلال منشوراتهم، مما قد يكشف عن حالاتهم النفسية. يمكن لتحليل النصوص في المنشورات أو استخدام تقنيات التعلم العميق أن يكشف عن .مشاعر وأحاسيس مستخدمي هذه المنصة

تمتد أهمية هذا البحث إلى صانعي القرار، واللجان التوجيهية، والمعلقين، والشركات الكبرى، إذ يُمكن أن يُساعد في إدارة شبكات التواصل الاجتماعي وفهم وتحليل الجوانب العاطفية. بالإضافة إلى ذلك، تُلقي الدراسة الضوء على المشكلات النفسية التي قد يواجها بعض مستخدمي وسائل التواصل .الاجتماعي، مثل التنمر الإلكتروني، والاعتلال النفسي، والنرجسية

ختامًا، تُساعد هذه الدراسة على فهم سلوكيات المستخدمين على منصات التواصل الاجتماعي مثل تويتر. كما يُمكن أن تُساعد نتائج هذه الدراسة في فهم المشاعر، وتعزيز الوعي بالصحة النفسية، وإدارة .محتوى وسائل التواصل الاجتماعي