Analyzing Social Media Impact on Gen Z and Gen Alpha Well-being Using Machine Learning and NLP Techniques

By
Mahmuda Islam
ID: 213-15-4464

FINAL YEAR DESIGN PROJECT REPORT

This Report Presented in Partial Fulfillment of the Requirements for the **Degree of Bachelor of Science in Computer Science and Engineering**

Supervised by

Tapasy Rabeya

Lecturer (Sr.Scale)
Department of Computer Science and
Engineering Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY
Dhaka, Bangladesh

September 16, 2025

APPROVAL

This Project titled **"Analyzing Social Media Impact on Gen Z and Gen Alpha Well-being Using Machine Learning and NLP Techniques** submitted by **Mahmuda Islam**, ID No: **213-15-4464** to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on **16 September, 2025.**

BOARD OF EXAMINERS

Dr. Sheak Rashed Haider Noori — **Chairman**
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Md. Sadekur Rahman — **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Mr. Saiful Islam — **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Dr. Md. Arshad Ali — **External Examiner**
Professor
Department of Computer Science and Engineering
Hajee Mohammad Danesh Science & Technology University

DECLARATION

We hereby declare that this project has been done by us under the supervision of Tapasy Rabeya, Lecturer (Sr. Scale), Department of Computer Science and Engineering, Daffodil International University. We also declare that neither this projectnor any part of this project has been submitted elsewhere for the award of any degree or diploma.

Supervised by:

Tapasy Rabeya
Lecturer (Sr. Scale)
Department of Computer Science and Engineering
Daffodil International University

Submitted by:

Mahmuda

Mahmuda Islam
Student ID: 213-15-4464
Department of Computer Science and Engineering
Daffodil International University

ACKNOWLEDGEMENTS

This work would not have been possible without the support and contributions of many individuals over the past two semesters. We are deeply grateful to everyone who has assisted us in one way or another.

First, we express our heartfelt thanks and gratefulness to the almighty for His divine blessing making it possible for us to complete the **Final Year Design Project (FYDP)** successfully.

We are grateful and wish our profound indebtedness to **Tapasy Rabeya**, **Sr. Lecturer**, Department of Computer Science and Engineering, Daffodil International University, Dhaka, Bangladesh. Deep knowledge and keen interest of our supervisor in the field of **Machine Learning and NLP** to carry out this project. Her endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts, and correcting them at all stages have made it possible to complete this project.

We would like to express our heartfelt gratitude to the Head of the Department of Computer Science and Engineering, for his kind help in finishing our project and also to other faculty members and the staff of the Department of Computer Science and Engineering, Daffodil International University.

We would like to thank our entire course-mates at Daffodil International University, who took part in this discussion while completing the coursework.

Finally, we must pay due respect to our parents for their continued support and patience.

ABSTRACT

The rapid advancement of the digital world has greatly changed how younger generation communicate, learn, and see their place in society. As Generation Z grows up and Generation Alpha enters school, it is important to understand how social media use is affecting their mental health. This study looked at how online activity affects behavior, emotions, and relationships. The results suggest that social media can have both positive and negative effects on self-esteem and stress. Those who use digital tools correctly are more confident and better problem-solvers. Social media helps people stay connected, but it also makes it harder to control emotions and increases the tendency to make unhealthy comparisons with others. The study also suggests that it has an impact on the cultural and social life of Generation Alpha in Bangladesh, which highlights the need for proper education on healthy technology use.

Table of Contents

Approval i

Declaration ii

Acknowledgements iii

Abstract iv

List of Figures vii

List of Tables viii

1 Introduction 1

1.1 Introduction 1
1.2 Motivation 2
1.3 Objectives 3
1.4 Methodology 3
1.5 Project Outcome 4
1.6 Organization of the Report 5

2 Background 7

2.1 Introduction 7
2.2 Literature Review 8
2.3 Gap Analysis 11
2.3.1 Intended contribution 12
2.4 Summary 14

3 Research Methodology 15

3.1 Methodology 15
3.1.1 Overview 15
3.1.2 Proposed Methodology/ System Design 16
3.1.2.1 Data Collection 17
3.1.2.2 Data Preprocessing 19
3.1.2.3 Feature Engineering 20
3.1.2.4 Model Training 21
3.1.2.5 Model Evaluation and validation 23
3.1.2.6 Analysis Confusion Matrix 25

3.1.3 Visualization Interface components 27
3.2 Detailed Methodology and Design 27
3.2.1 Alternative Solutions Considered..27
3.2.1.1 Text Preprocessing Alternatives 27
3.2.1.2 Machine Learning Alternatives 28
3.1.2.3 data Collection Methodology Alternatives 29
3.2.2 Selected Solution justification 29
3.2.2.1 Methodology Selection Rationale 29
3.2.2.2 Technical Implementation Justification 30
3.3 Project Plan 31
3.3.1 Project Timeline and Milestone 31
3.3.2 Risk Management and Backup Plans 31
3.3.3 Quality Assurance Framework 32
3.4 Task Allocation.. 34
3.5 Summary 34

4 Implementation and Results **35**
4.1 Environment Setup 35
4.2 Testing and Evaluation/Performance/ Comparative Analysis 35
4.3 Visualization Results 38
4.4 Results and Discussion 41
4.5 Summary 42

5 Engineering Standards and Design Challenges **43**
5.1 Compliance with the Standards 43
5.1.1 Software Standards 43
5.1.2 Hardware Standards 44
5.1.3 Communication Standards 44
5.2 Impact on Society, Environment and Sustainability 45
5.2.1 Impact on Life 45
5.2.2 Impact on Society & Environment 45
5.2.3 Ethical Aspects 46
5.2.4 Sustainability Plan 46
5.3 Project Management and Financial Analysis 47
5.4 Complex Engineering Problem 48
5.4.1 Complex Problem Solving 49
5.4.2 Engineering Activities 49
5.5 Summary 50

6 Conclusion **51**
6.1 Summary 51
6.2 Limitation..51
6.3 Future Work..52

References 54

List of Figures

3.1	Flowchart of Proposed Methodology	16
3.2	Sample Dataset	18
4.1	Bar Chart for Model Comparison	36
4.2	Confusion Matrix of Target column 1 and Target Column 2	36
4.3	Pie Chart of Target Columns	38
4.4	Bar Chart for Target Column 2	38
4.5	Pair Plot of Key Features	39
4.6	Processed Text Overview	39

List of Tables

2.1 Summary of Literature Review 8

2.2 Comparative Analysis of Selected Studies and This Research 12

3.1 Survey Questions by Category 18

3.2 Task Allocation Table 32

4.1 Per Model Performances Summary 35

4.2 Per Model Performances Summary (Macro and Weighted Averages) 35

5.1 Sustainability Planning Table 46

5.2 Estimated Cost Table 47

5.3 Mapping with Complex Engineering Problem. 48

5.4 Mapping with knowledge Profile 48

5.5 Mapping with Complex Engineering Activities 48

Chapter 1

Introduction

This chapter explains the background, purpose, goals, and importance of research on the impact of social media on the mental health of secondary school students in Bangladesh.

1.1 Introduction

Today, being online has become a big part of how people communicate and engage with society. Teenagers are at the center of this change, where the way people grow and develop is changing. This age is crucial for identity formation, social skills acquisition, and emotional development. But today's teenagers are facing these challenges in a digital environment more than ever before. Students aged 13 to 19 (grades 8 to 11) spend an average of 5 to 6 hours or more per day on social media and various online platforms, which is much more than their school time. [1]. Globally, more than 93% of teenagers go online and at least 46% of them are active on social media, some for several hours a day. Teenagers use social media platforms, such as Facebook, Instagram, TikTok, Snapchat and Twitter for an average of 3.5 hours daily. [2][8].

The time teenagers spend with digital devices and social media has increased a lot since the Covid-19 pandemic began. This is because of remote learning, less face-to-face time with friends and less time outdoors [10]. In Bangladesh, especially research shows that 88.25% of school-going teens had internet addiction during the pandemic, and 72.51% felt lonely. This has made emotional and psychological challenges worse for these young people, especially as the move from one generation to the next [3].

The effects of being deeply involved with digital platforms go beyond just leisure. Data from around the world shows that children and teens who spend over 3 hours a day on social media are twice as likely to have mental health issues, such as depression and anxiety, according to WHO Europe, the number of teens with problematic behavior. Research also shows that 67% of students aged 13 to 19 feel anxious when they can't use their devices for a while, and 54% check their social media during class [9]. Even more concerning, 43% of this age group have experienced cyberbullying, and 61% compare themselves to others online, which hurts their mental health and school performance [6].

The relationship between social media use and well-being in school-aged teens is not entirely clear. 70% of young people concede that social media occasionally makes them feel bad or anxious, but 74% say these platforms are crucial for emotional support and maintaining contact with friends [5]. This shows that social media, like any other medium, has both good and bad sides. It can bring people together and give them a chance to express themselves, but it can also make some people feel ignored or constantly judged.

Being able to understand and judge online information is one of the main ways that social media impacts people's lives. [4]. However, the data shows that only 31% of students in grade 8 through 11 feel confident in their ability to analyze online information, and even fewer – around 19% - have received any formal education about healthy digital behavior. This difference is even greater for Generation Alpha, as they have been using smartphones since childhood and their cultural influence and mental maturity are different and less than those of Generation Z. [8].

With concerns about the mental health of adolescents around the world increasing, it is now imperative to understand how social media is affecting them [10]. This study analyzed the impact of social media on secondary school students in Bangladesh. It used survey data, natural language processing (NLP), and machine learning (ML) to explore the psychological impacts of online activities.

Ultimately, this research can help develop educational programs and digital well-being plans to help teens use social media in a healthier and more positive way. It also addresses the specific challenges of the generation that is moving from digital adoption (Generation Z) to digital normality (Generation Alpha).

1.2 Motivation

In today's fast-paced world, teenagers are spending a lot of time on social media, not just in school. This can have an impact on their mental health and communication skills. Around 93% of teenagers worldwide use social media, and students in Bangladesh spend around 5–6 hours a day on various digital platforms. This shows how important it is to understand their online activities and their impact on their mental health. According to the latest data from WHO Europe, the number of people experiencing problems with social media use increased from 7% in 2018 to 11% in 2022, which highlights the need for a new computational approach to analyzing this issue.

This study is based on alarming data on the mental health of adolescents in Bangladesh. During the COVID-19 pandemic, 82% of students reported internet addiction, and 72.51% of students said they felt lonely (Bolle et al., 2020). Traditional mental health assessment methods are not able to fully measure the complex emotions associated with students' online lives and social media use. To fill this gap, natural language processing (NLP) can be used to detect emotional signals and sentiment patterns that may not be seen in general surveys, and this allows for a better understanding of the relationship between social media and mental health.

Additionally, schools in Bangladesh do not have sufficient systems in place to teach digital wellbeing or safe online use. Only 31% of students are confident in judging online content. This shows that computer-based tools are needed to create effective solutions. Using advanced data science and machine learning methods, this project aims to create scientifically based digital wellbeing models that can be applied across cultural contexts. Studying the differences between Generation Z and Alpha will also help improve educational technology and mental health programs.

Addressing this issue is important not only for researchers, but also for policymakers, mental health professionals, and technology developers. They need robust data to develop tools and programs to support digital well-being. The computational methods developed by this study can be used for more extensive studies of how teenagers interact with digital platforms, extending our understanding in a global perspective about how new technology shapes mental development in different cultures.

1.3 Objectives

- Explores the impact of social media on the mentality of Generation Z and Alpha students.
- Using NLP algorithms and Machine Learning, extract and classify emotional trends.
- Identify how generations cope with virtually and emotionally.
- Develop evidence-based methods and frameworks for digital wellness.
- Examine the partnership between media literacy and psychological resilience.
- Design tools for automatically analyzing sentiment.
- Increase academic discussions related computational psychology.
- Offer practical solutions to schools, legislators and parents.

1.4 Methodology

This study is a Hybrid qualitative-analysis with traditional survey approaches combined with advanced computational tools to understand the psychological impact of social media on Bangladeshi secondary school students who are 13-19 years old . A total of 231 responses were gathered through Google Forms, employing a cross-sectional design to capture a snapshot of emotional trends, behavioral characteristics, and digital engagement among Generation Z and Generation Alpha students. The survey consists of structured demographic queries as well as open-ended text responses, facilitating an in-depth structured analysis through Natural Language Processing (NLP).

For the analysis of qualitative text submissions, the research employs a custom text preprocessing pipeline featuring standard techniques such as converting to lowercase, removing punctuation, filtering out stopwords, and TF-IDF vectorization. These approaches convert raw student narratives into numerical feature vectors, which are subsequently analyzed using four different machine learning algorithms: Logistic Regression, Naïve Bayes, Random Forest, and Support Vector Machine (SVM). The dataset was divided with an 80:20 train-test ratio, and a 10-fold cross-validation was conducted to ensure a thorough evaluation of classification accuracy.

1.5 Project Outcome

This is not just a technical task. It also has social, educational, and scientific aspects:

Sentiment Analysis: The system should accurately interpret students' open-ended responses—whether positive, negative, or neutral. We can also understand how middle school teens engage emotionally on social media, such as by detecting patterns of stress, anxiety, excitement, and happiness.

Model Performance Evaluation: Four machine learning methods—SVM, Random Forest, Naïve Bayes, and Logistic Regression—were tested to see which ones were most accurate and reliable for small educational surveys. These results may help future research when working with similar limited data.

Interactive Dashboard & Data Visualization: A simple dashboard has been created that displays graphical data such as bar charts, pie charts, confusion matrices, and word clouds. These visuals allow students, teachers, or policymakers to quickly understand complex emotional information and instantly see how everyone is feeling.

Academic Contribution: This study provides a clear methodology, which will help future research understand adolescents' mental and digital well-being.

Recommendations: This project provides a clear and reusable approach to understanding the psychological impact of social media. Because it is designed in a circular manner, other researchers can use, modify, or extend it across cultures or communities.

Well-being Suggestions: This article offers some ideas for helping students feel happier online and offline. These ideas can help teachers, school programs, and mental health activities in Bangladesh, and can be used in schools in other countries as well.

Scalable and Flexible Framework: Although this study only received 500 responses, the system can handle many more. It is designed to make it easy to add new data in the future or to match it with data from other schools and researchers.

Ethical Standards: This project-maintained confidentiality and ethics. Students remained anonymous, their information was secure, and everyone knew what would happen so they participated. This builds trust and will help inform larger studies in the future.

Multidisciplinary Impact: This research connects computer science, psychology, and education. It shows how technology can be used to understand adolescents' mental health and online experiences.

Guidance for policy and Institutional Action: The study provides clear and actionable advice for educators and policymakers about how social media impacts students' learning, emotions, and health.

Foundation for Future Longitudinal Research: Although this study used a one-time survey, the results could inform future research that looks at changes in students' emotions and well-being over time.

1.6 Organization of the Report

This report consists of six interconnected chapters, which provide important insights into the policy and research context:

Chapter 1: Introduction – The chapter describes the background, objectives, goals, and other related issues of the research. It also explains what the research project covers, how it was done, for whom, and what its main goal is.

Chapter 2: Background and Literature Review – The report reviews relevant literature on Generation Z and Generation Alpha, analyzes how social media is influencing emotional and social trends, and how they are using digital tools. It also critically reviews existing research on technology use, mental health, and sentiment analysis, and explains how the current study was selected from previous studies for the Bangladeshi context.

Chapter 3: Research Methodology – This section describes the research methodology. It explains the survey design, participant selection process, and data collection methods. The text processing procedures, feature extraction using TF-IDF, and the application of four machine learning models (SVM, Random Forest, Naïve Bayes, and Logistic Regression) are then covered in detail. It also describes the rationale behind the selection of particular techniques, the alternatives that were taken into account, and the design, management, and risk assessment of the system.

Chapter 4: Implementation and Results – This chapter explains how the sentiment analysis system was built and tested. It shows how the data was managed, the model trained, and the accuracy was verified. Visual tools like word clouds and graphs made it easier for students to understand emotions.

Chapter 5: Engineering Standards & Societal Impact – This chapter assesses the technical standards and practical impacts of the project. It discusses ethical engineering practices, environmental and ethical issues, and how technology can enhance integrity and sustainable learning.

Chapter 6: Conclusion and Future Work – The final chapter highlights key findings and assesses the extent to which the project has achieved its goals. It highlights current limitations and potential for future development, while clarifying the research's importance to educators and those interested in student well-being and learning through digital media.

Chapter 2

Background

This chapter revies existing research on social media usage, generational psychology, sentiment analysis, and machine learning applications in mental health studies.

2.1 Introduction

The continuing digital revolution in Bangladesh has significantly transformed how young people socialize, learn, and understand their place in society. With internet penetration rate reaching almost 72% across the country and almost 89% of adolescents owning smartphones, social media becomes essential to the daily activities of Bangladeshi students. This change is evident especially in urban areas, where 94% of grades 8-11 students actively engage with sites like Facebook, TikTok, Instagram, Snapchat, X (Twitter), and YouTube. Such widespread online engagement confirms the importance of social networking sites in communicating and building identity, entertainment, and peer relationships among young people.

The widespread youth engagement in social media intersects with troubling trends in mental health. Current studies by the Bangladesh Association of Mental Health reveal that 42% of teenagers at the secondary school level show anxiety symptoms, while 36% are depressed or constantly sad. Moreover, close to 58% of 12- to 19-year-olds say they employ unhealthy social comparison on social media, often comparing their lives to perfected online displays that establish unattainable standards of success, appearance, and happiness. These behaviors are reflective of social comparison theory, in that adolescents are more likely to hold negative opinions regarding themselves when they repeatedly expose themselves to idealized images of their peers and role models.

The local cultural and socio-economic environment in Bangladesh further exacerbates these issues worldwide. High levels of societal pressure to perform well academically lead to stress; 76% of students perceive that they must show their academic achievement on social media. This virtual self-promotion serves as a way to hide real problems; students try to present themselves as confident and successful. But poor students are even more anxious because they want to meet these online expectations, while in reality they are suffering. A 2023 study conducted by BRAC University found that 63% of students from middle-class families struggle with the finances to maintain their online lifestyle, such as buying the latest smartphone or trendy clothes. This creates a vicious cycle, where online recognition becomes detrimental to financial well-being and mental health.

Evidence on the impact of social media shows that it creates a complex mix of opportunities and risks, consistent with the experiences of adolescents in Bangladesh. Social media increases communication with relatives abroad, provides access to digital learning resources, and exposes students to diverse global perspectives—benefits not previously available to many Bangladeshi students. Yet, alongside these benefits, there are also significant risks that disproportionately affect young users. These negative effects include stress arising from social comparison, poor media literacy, gender differences in online communication, and increased sensitivity to the influence of online peers. Taken together, these findings highlight the importance of adapting international social media research to the specific cultural, social, and economic context of adolescents in Bangladesh. This recognizes both the benefits and limitations of online participation.

2.2 Literature Review

The following is an extensive summary of current research on social media use and psychological well-being among Generation Z and Generation Alpha. From the global and regional points of view, the literature provides a thorough explanation of how online environments contribute to mental health, identity formation, and social connections.

Current studies indicate concern that social media use by younger generations is irrevocably intertwined with positive and negative psychological effects. For example, Xuan et al. (2025) and Piccerillo et al. (2025) illustrate how smartphone addiction and early exposure to looks-oriented content correlate directly with heightened stress, lower emotional intelligence, and vulnerability to sociocultural pressures. These findings suggest that technology does provide access, but that it also perpetuates issues of body image and performance expectation.

Conversely, research also highlights online support features. Doung et al. (2025) and Khalaf (2023) mention that social media could be used to enable bonding, cultural integration, and emotional support, especially among expatriates and youth who are seeking acceptance. However, the same media also bring dangers such as cyberbullying, fear of missing out (FOMO), and unfavorable social comparison, which could be more prominent where media literacy levels are low.

Growing evidence is calling attention to the psychological flexibility needed to counter these risks. Lim et al. (2024) and Sharma et al. (2024) assume that psychological flexibility, self-acceptance, and new media literacy can act as buffers for adolescents. These can help them respond to online pressure and regulate emotional reactions. They recommend intervention strategies like resilience-building education programs that develop critical thinking skills in online contexts.

Regional research, particularly in the South Asian context, provides further insight. Irfan et al. (2022) noted that excessive screen use among adolescents in the region increases anxiety, reduces attention span, and causes overall emotional instability. Similarly, Ziatdinov and Cilliers (2022) addressed the educational aspect, noting that the technology-aware learning style of Generation Alpha requires strong educational models that embrace technology, rather than against it.

Overall, the literature considered here shows that the impact of social media on adolescents' mental health is not simply positive or negative. Rather, it is the result of a complex interaction between user usage patterns, social and cultural environments, and levels of digital literacy. The evidence is clear that interventions need to be culturally relevant and context-specific, especially in Bangladesh where educational aspirations and socio-economic disparities increase uncertainty about digital well-being outcomes.

Author (s)	Year	Title	Methodology	Key Findings
Wei Xuan, Meghna Roy Chowdhury, Yi Ding, Yixue Zhao	2025	Unlocking Mental Health: Exploring College Students' Well-being through Smartphone Behaviors	Quantitative Analysis, data-driven study.	Findings reveal significant variations in smartphone usages across gender and location, highlighting links to mental health and offering opportunities for predictive models to enhance digital well-being.
Rushan Ziaatdinov, Juanee Cilliers	2022	Generation Alpha: Understanding the Next Cohort of University Students	Theoretical analysis	Poised generation Alpha to reshape higher education with their tech-savvy, socially connected, and highly perceptive learning preferences, demanding innovative and personalized teaching approaches.
Abderrahman M. Khalaf	2023	The Impact of Social media on Mental Health of Young Adults: A Literature review	Qualitative Analysis	Showed the connection and source of psychological distress among young adults, with negative effects like anxiety, depression, and low self-esteem often driven by social comparison, cyberbullying, and the pressure to maintain idealized online personas.
Nadine Agyapong-Opoku, Felix Agyapong-Opoku, Andrew J. Greenshaw	2025	Effects of Social Media Use on Youth and Adolescent Menta; Health: A Scoping Review of Reviews	Qualitative analysis	Showcased the impact on adolescent mental health is multifaceted and highly individualized, with problematic and passive use linked to depression and anxiety, while active engagement can foster social support and reduce isolation.

Dr. Nidhi Sharma, Dr. Raksha Thakur, Hasan Hamid, Kunal Sadwani	2024	The Perils of Social Media: Its Impact on Youth Mental Health, Marital Relationships, and Effect on Generation Alpha	Survey study	Research has shown that excessive social media use among adolescents leads to negative mental health and relationship dissatisfaction, which later impacts generations, especially Generation Alpha.
Lidia Piccerillo, Alessia Tescione, Alice lannaccone, Simone Digennaro	2025	Alpha Generation's Social Media Use: Sociocultural Influences and Emotional Intelligence	Quantitative analysis	Predicted addiction, reinforces sociocultural appearance ideals, and negatively impacts emotional intelligence on social media usage among Generation Alpha
Yuehua Han, Zhifen Xu	2024	Fostering College Students' Mental Well-being: The Impact of Social Networking Site Utilization on Emotion management and regulation		
Syed Bilal Irfan, Malavika M, Aaminah Firdos. Afifa Sadaf	2022	A study Examining the Impact of Social Media Use on Cen-Z Well-being	Quantitative analysis	Overuse of social media among Gen Z correlated with increased anxiety, reduced attention span, and deteriorating mental and physical well-being.
Jia Hang Lim, Mahdir Bin Ahmed, Kususanto Ditto Prihadi	2024	New Media Literacy, Self-Acceptance, and Psycological Flexibility in Enhanching Gen Z Well-being	Quantitative analysis	Psychological flexibility significantly enhances Gen Z's well -being, especially when medicted by self-acceptance and supported by new media lieracy.

Hoai Lan Duong, Thi Kim Oanh, Minh Tung Tran, Thi Kim Cuc Tran	2025	Social Media's Role in enhancing Psychological Well-being of Generation Z Expatriates	Qualitative analysis	Social media strengthens psycological well-being among Gen Z expatriates by fostering virtual connections, cultural integration, and social support across borders.

Table 2.1: Summary of Literature Review

2.3 Gap Analysis

The existing literature on the relationship between social media and mental health has revealed several important gaps that need to be explored further. One major issue is the heavy reliance on self-report survey data, which is a common but problematic approach. While such data can provide general insights, they are often superficial and not reliable for accurately identifying real behavioral trends. This over-dependence weakens the reliability of study findings, as self-reports can be affected by memory mistakes, the desire to look good, and differences in how individuals interpret their experiences.

Another important gap involves Generation Alpha, who engage with digital technologies at an earlier age than Generation Z. So far, there has been very little research on this generation, and even less on the long-term impacts of early and frequent digital exposure on development and psychological well-being. Although more attention is now being given to how childhood and adolescent technology use influences future social skills, identity, and emotional resilience, few studies include reliable data or predictive models to support these ideas.

Some studies also discuss media literacy, psychological flexibility, and emotional resilience, but they often ignore the differences in people's psychological adaptations and the role of cultural context. This is especially true in South Asia, where cultural beliefs such as educational attainment, family status, and economic inequality greatly influence digital behavior. Ignoring these cultural factors leads to generalizations in current research that may not always apply to different populations.

Furthermore, the complex connections between digital literacy, emotional intelligence, and emotional resilience have not been well explored. Few research models integrate these components to examine how adolescents cope with stress arising from online activities. Without this connection, interventions are often untargeted and less effective in addressing the underlying causes of stress associated with Internet use.

Finally, there is a lack of longitudinal and mixed-methods research in this area. Most previous research has used **cross-sectional designs**, which provide only a snapshot of digital behavior. This approach does not show how usage patterns and their psychological impact change over time. Furthermore, the limited use of mixed-methods reduces the depth of the findings, as quantitative data alone cannot fully capture social, cultural, and emotional components. Addressing these methodological issues is essential to more fully understand the psychological impact of digital communication across generations and cultures.

2.3.1 Intended Contribution

My study attempts to bridge some very important gaps by conducting an extensive and relevant analysis of the impact of social media on teenagers. It narrowly focuses on Generation Alpha and Generation Z in South Asia, paying specific attention in the case of Bangladesh. The region is experiencing rapid digitization along with special cultural, social, and economic factors, making it a case of special interest for studying the impact of digital media. The study examines the effects of first exposure to new media and differences in media literacy on emotional and cognitive development. The study also seeks to investigate emerging tendencies normally overlooked by research focusing on Western populations.

For this purpose, the study used a mixed-methods approach, combining quantitative measures of adolescents' user activities with qualitative feedback. This two-step approach identifies clear patterns of social media activity and subjective feelings of users' online experiences, providing a holistic understanding of digital communication. It is sensitive to the shadow effects of online activity and its emotional responses.

An important aspect of the study is psychological flexibility and emotional intelligence, which refer to how social media use affects adolescents' psychological well-being. Analyzing these psychological traits can help explain why some adolescents can cope with adversity and remain psychologically resilient in online situations, while others suffer from problems such as depression, anxiety, and negative self-esteem. The focus on individual differences raises questions that are often unexplored in general or international research.

In addition, this study has developed a contextual framework that takes into account the Bangladeshi context. It includes factors such as socio-economic class, family circumstances, and educational stress, which influence online behavior, but are often omitted in internationally conducted research. This context-sensitive framework ensures that the research findings are also valid in local and homogeneous contexts, which can be applied to similar cultural environments.

Overall, this study helps to bridge generational and regional gaps and increases the methodological value of the topic. By combining context-sensitive discovery and formal machine learning methods, it provides a re-applicable framework that can be tested in future digital wellbeing research. Thus, the paper offers an applied template that offers evidence-based guidelines for instructors, policymakers, and mental health experts interested in building more supportive digital environments for youth.

Feature/Focus Area	Wei et al. (2025) – Smartphone Behaviors	Piccerillo et al. (2025) – Gen Alpha & Emotional Intelligence	Khalaf (2023) – Social Media & Young Adults	Sharma et al. (2024) – Youth & Mental Health	Irfan et al. (2022) – Gen Z Well-being	This Study (Bangladesh, 2025)
Generation Studied	Gen Z (college students)	Gen Alpha	Young Adults	Gen Z & Gen Alpha	Gen Z	Gen Z & Gen Alpha
Geographical Focus	Global	Europe	Middle East	India	South Asia	Bangladesh
Key Constructs	Smartphone usage & stress	Emotional intelligence, appearance ideals	Anxiety, depression, cyberbullying	Mental health, social media addiction	Screen time, anxiety, attention	Sentiment, psychological impact
Methodology	Quantitative	Quantitative	Qualitative review	Survey-based	Quantitative	Mixed (survey + ML)
Psychological Factors	Stress, well-being	Emotional intelligence	Self-esteem, anxiety	Depression, marital/social impact	Anxiety, well-being	Emotional intelligence, psychological flexibility
Cultural/Socioeconomic Context	Limited	Sociocultural focus	Limited	Cultural expectations in India	Regional context	Strong focus (Bangladesh, academic, and socioeconomic pressure)
Machine Learning/NLP	X	X	X	X	X	√ (TF-IDF, ML models)
Visualization/ Dashboard	X	X	X	X	X	√
Ethical Consideration	X	X	X	X	X	√ (Consent, anonymization)

Table 2.2: Comparative Analysis of Selected Studies and This Research

As illustrated in Table 2.2, prior research has explored the connection between social media and mental health among young people in various regions and groups. Nevertheless, many of these studies did not effectively combine different research methods or keep pace with technological advancements. For instance, Wei et al. (2025) and Irfan et al. (2022) primarily used quantitative approaches to study smartphone use and screen time, but they did not incorporate more sophisticated computational tools. Likewise, Piccerillo et al. (2025) and Sharma et al. (2024) investigated emotional intelligence and psychological impacts without sufficient consideration of cultural factors. Khalaf (2023) provided a comprehensive summary of the existing literature, but it lacked empirical evidence and technical depth.

2.4 Summary

The rapid spread of social media has fundamentally changed the way young generations interact, learn, and express themselves. For Generation Z and the new Generation Alpha, digital platforms are not just communication tools, but also a critical environment for identity formation, emotion regulation, and social acceptance. These platforms provide unprecedented access to information and peer networks, but they also pose complex psychological challenges, such as anxiety, depression, and attention deficit disorder.

In particular, the South Asian context creates a unique situation where cultural expectations, academic pressures, and group values intersect with digital behavior. Adolescents and young adults in regions like Bangladesh often use social media within close-knit family structures and high-pressure educational institutions, which exacerbates the emotional impact of online comparison, cyberbullying, and fear of missing out (FOMO).

Despite growing global interest in the mental health impacts of digital life, existing research has generally focused on populations in Western countries, leaving a gap in understanding how sociocultural factors influence digital experiences in non-Western contexts. In addition, research on Generation Alpha is limited, particularly on how their early and deeper digital platform experiences impact emotional intelligence and cognitive development.

Chapter 3

Research Methodology

This chapter presents the methodology used to study the psychological impact of social media on secondary school students in Bangladesh. This methodology combines conventional survey methods and advanced computational technology to provide a comprehensive analysis.

3.1 Methodology

3.1.1 Overview

This study used machine learning and computational linguistics to examine the impact of social media on the mental well-being of Generation Z and Generation Alpha students in Bangladesh. To ensure comprehensive analysis, data were collected using specific criteria and various methods of assessing mental skills, including open-ended and closed-ended questions. The selection of students was done in a way that ensured a diversity of academic fields, educational backgrounds, and socio-economic levels, thus ensuring reliable and diverse results.

The study analyzed handbooks written in Bengali, English, and other languages using natural language processing (NLP). Behavioral trends and emotional expressions were also identified based on the answers given by the students. Sentiment analysis used conventional machine learning methods, such as Support Vector Machine (SVM), Random Forest, Naive Bayes, and Logistic Regression.

The main goals of the process during the data management, model training and evaluation stages are accuracy, fairness and effectiveness. Extensive reports and graphical dashboards are used to display the results. Considering scalability, security and sustainability, the framework is designed for long-term use, suitable for both education and health sectors.

As a result, this system provides a robust, modern and culturally relevant way to understand the psychological impact of social media on adolescents in Bangladesh.

Research Methodology Components:

i. Cross-sectional survey design targeting secondary school students aged 13-19

ii. Stratified convenience sampling across diverse educational institutions in Bangladesh.

iii. Validates psychological assessment instruments integrated with open-ended response collection.

iv. Statistical analysis framework incorporating descriptive and inferential statistical methods.

Requirements:

i. Scalable pipeline for text pre-processing methods supporting multilingual content analysis.

ii. TF-IDF vectorization module for feature extraction from textual responses.

iii. Multi-algorithm machine learning framework supporting four distinct classification approaches.

iv. Comprehensive evaluation metrics system with cross-validation capabilities.

v. Interactive visualization dashboard for performance comparison and result interpretation.

3.1.2 Proposed Methodology

This diagram (fig: 3.1) shows how a machine learning experiment using Natural Language Processing (NLP) is completed. The process begins with data collection; then follows the steps of model building, validation, and characterization.

Gathering and Preprocessing of Data: Data collecting, the first phase, involves questionnaire data gathering using Google Forms. Once the data has been gathered, the preprocessing process starts. This step is crucial as it gets the raw textual information ready for machine learning algorithms. The procedures entail eliminating punctuation and numbers, familiar words that are not particularly significant (like the or), and converting all words to lowercase. Using these procedures to clean up the data may make the model much easier to comprehend.

Feature Engineering and Model Construction: The next step, feature engineering, transforms the cleaned text into one usable by machine learning models. This method, as shown by the flowchart, makes use of TFIDF vectorizing. This method converts words into figures according to their relevance in a text and throughout the whole database. The information is then divided into two ways: one for model training, and another for testing. This gives the model confidence in its capacity to manage new, previously unknown data. Following this outcome, the model training stage during which multiple kinds of models – including Logistic Regression, Naïve Bayes, Random Forest, and SVM – are trained using the previously processed data.

Assessment and analysis of models: After training, the models are thoroughly examined repeatedly. The flowchart depicts three distinct evaluation stages, therefore suggesting a thorough approach for assessing the models' performance utilizing metrics such as Precision, Recall, and F1score. Cross validation guarantees the findings are consistent. It is therefore also included. Following the evaluations, a review is carried out. This comprises establishing word clouds to reflect the most often used words as well as a confusion matrix to pinpoint the regions where the model made faults; comparing several models to pick the best one. Interpreting and presenting the results, the last step clarifies the study's and assessment's consequences.

Here's a flowchart in figure 3.1 to represent proposed methodology:

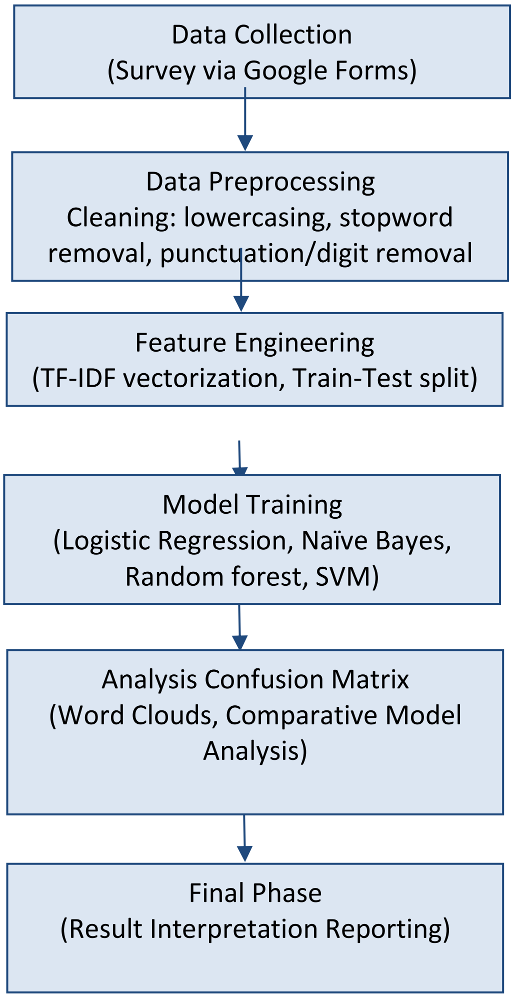

Figure 3.1: Flowchart of Proposed Methodology

3.1.2.1 Data Collection

The data collection process in this research combined structured surveys with psychological assessment scales to obtain quantitative as well as qualitative data about students' mental health and social media usage. The survey was conducted to have a mixture of demographic questions, social media behavior measures, validated psychological scales, and open-ended questions. To ensure the representativeness, a stratified sampling method of multiple stages was employed in each of the eight divisions of Bangladesh, including students from both urban and rural colleges, different socioeconomic classes, and differing education systems (government schools, government colleges, private colleges, and private schools).

The survey was done in offline and online modes. Online surveys were posted on education websites and social media (Google Forms), while offline surveys were done in institutions with weak digital connectivity. All participants provided consent, and the answers were anonymized for protection.

The survey instrument included:

i. Demographic Information: Age, Gender, Level of education.

ii. Use of Social Media: Time spent on social media, Preferred platform, Posting frequency, Engagement behavior.

iii. Psychological Scales: Revised forms of Depression Anxiety Stress Scales (DASS-21)

Open-ended Questions to capture subjective thoughts regarding social media usage, mood states, and perceived academic/relationship impacts.

Category	**Survey Questions**
Demographics	Gender, Class, Age
School & Study Time	On a typical weekday, how much time do you spend outside your home for school + private (or coaching)-related activities (including travel to/from school and time spent in school)?
Social Media Usage	On a typical day, how much time do you spend on social media platforms (like Instagram, Facebook, YouTube, TikTok, Snapchat, X, etc.)?
	Which social media platforms do you use most frequently? (Select all that apply)
	What types of content do you typically watch or search for on social media? (Select all that apply)
Emotional Reactions	How do you feel after seeing any content?
	Do you feel that social media has a negative or positive impact on your self-esteem?
Psychological Well-being	Do you feel you have enough time for your hobbies and interests?
	Do you feel connected to your friends and family?
	General feelings: Do you ever feel depressed or sad after seeing other people's posts online?
Social Media Addiction & (FOMO)	Do you check again and again how many reacts or likes you get after posting something on social media?
	Are you concerned about the number of reacts and comments you get on your post?
Cyberbullying & Online Harassment	Have you ever experienced cyberbullying or online harassment?
Sleep & Concentration	How often does social media use interfere with your sleep?
	Do you find it difficult to concentrate on schoolwork because of social media distractions?
Social Isolation/Connection	Do you feel more or less socially connected because of social media?
	Do you feel lonely, even when you are online?

Physical Health & Activity	Do you experience any eye strain or headaches, neck or back pain after using social media?
	How often do you participate in physical activity each week?
Open-ended	What strategies do you use to manage the negative effects of social media?
	Do you take breaks from social media?

Table 3.1: Survey Questions by Category

Here is a sample dataset in figure 3.2 of our research:

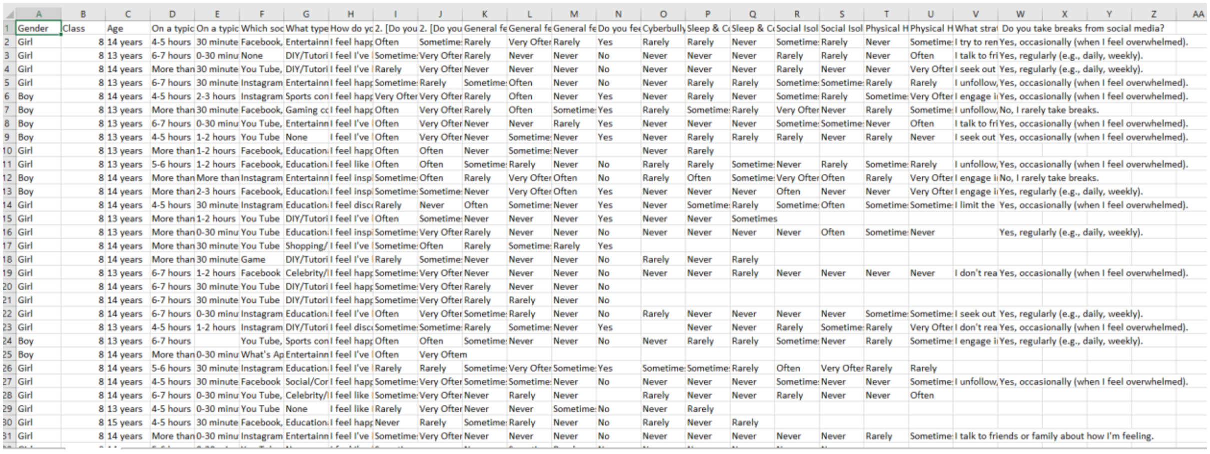

Figure 3.2: Sample dataset

3.1.2.2 Data Preprocessing

This research focused on preprocessing data for machine learning and natural language processing. The dataset included demographic details from surveys and unstructured text comments from Bangladeshi secondary school students. Before training the machine learning models, raw survey responses underwent a structured **Natural Language Processing (NLP) pipeline** to ensure that the data was both clean and suitable for numerical feature extraction. Since the survey responses included both English and Bengali text, standard preprocessing steps were applied consistently across languages to reduce noise and enhance model interpretability.

Text Cleaning and Normalization

1. **Lowercasing**: All text responses were converted to lowercase to standardize input.

Before Lowercasing	**After Lowercasing**
I am **Happy** today.	i am **happy** today.
Social Media is **Good**.	social media is **good**.
Feeling **Excited** but Tired.	feeling **excited** but tired.

For example, words like *Happy* and *happy*, *Good* and *good*, *Excited* and *excited* were treated as the same token. This avoids redundant feature representations in the vocabulary.

2. **Removing Numbers and Punctuation**: Survey responses often contained numbers, punctuation marks, or typing errors that did not contribute to sentiment meaning.

These were removed using **regular expressions (regex)**, ensuring that only meaningful words were retained for analysis.

Before Removing Numbers and Punctuation	**After Removing Numbers and Punctuation**
I slept at 2:am :(	i slept at am sad
I got 95% marks, feeling sooo happy!!!	i got marks feeling sooo happy
Social media is fun… but sometimes tiring.	social media is fun but sometimes tiring

3. **Stopword Removal**: Commonly used words (e.g., *the, is, and*, or their Bengali equivalents like এবং, এটি, যে) carry little to no sentiment value. A predefined stopword list was applied to filter them out, thereby improving the signal-to-noise ratio in the dataset.

Before Stopword Removal	**After Stopword Removal**
The social media use is **very** high.	social media use very high
I think **it is** stressful sometimes.	i think stressful sometimes
Friends **and** family keep me happy.	friends family keep me happy

4. **Tokenization**: Each response was broken down into individual words or tokens. Tokenization is an essential step because it allows text to be represented in a structured format, enabling frequency calculations and vectorization.

Before Tokenization	**After Tokenization**
"Social media makes me happy today"	[social,media,makes,me,happy,today]
"I feel anxious after scrolling TikTok"	[i,feel,anxious,after,scrolling,tiktok]
"Online classes are boring sometimes"	[online,classes,are,boring,sometimes]

3.1.2.3 Feature Engineering

After text cleaning and tokenization, the responses were transformed into numerical features using **Term Frequency–Inverse Document Frequency (TF-IDF) vectorization**. Unlike simple frequency counts, TF-IDF reduces the importance of very common words while highlighting rarer, more meaningful terms.
The formula is:

$$\textbf{TF-IDF}(t,d) = \textbf{TF}\ (t,d) \times \log\frac{N}{DF(t)} \ldots\ldots\ldots\ldots\ldots\ldots \textbf{(i)}$$

Where:

- TF(t,d)= frequency of term t in document d
- N = total number of documents in the dataset
- DF(t) = number of documents containing term t

Interpretation:

- Terms that occur frequently in a document but rarely across other documents receive a high TF-IDF weight.
- Common terms that appear in nearly every document receive a low weight.

TF-IDF was selected over more advanced embedding techniques (e.g., Word2Vec, GloVe, Transformers) for several reasons:

- **Simplicity and Interpretability:** TF-IDF weights are easy to calculate and interpret, making them suitable for academic research where transparency is important.

- **Effectiveness for Small Datasets:** With only 231 survey responses, advanced deep learning embeddings would risk overfitting. TF-IDF performs reliably in small to medium-sized corpora.
- **Computational Efficiency:** TF-IDF requires fewer computational resources compared to deep neural embeddings, which aligns with the project's performance and scalability constraints.
- **Well-Established in Literature:** TF-IDF is a widely accepted feature extraction method in text classification and sentiment analysis, ensuring methodological consistency with previous research studies.

3.1.2.4 Model Training

Four supervised machine learning algorithms were implemented using **scikit-learn** in Python to classify sentiment:

Logistic Regression (LR)

A linear model used for classification that estimates the probability of a sample belonging to a specific class using the logistic (sigmoid) function. LR is simple, interpretable, and effective when the relationship between features and the target variable is approximately linear. In this study, LR provided a baseline for performance comparison.

Mathematical Formulation:
For a given input feature vector xxx, the probability that it belongs to class y=1y = 1y=1 is:

$$P(y=1 \mid x) = \sigma(w^T x + b) = \frac{1}{1+e^{-(w^T x+b)}} \text{.................. (ii)}$$

Where:

- w = weight vector
- b = bias term
- σ(z) = sigmoid activation function

A decision boundary is set at $P(y=1|x) \geq 0$. Model parameters are estimated via **maximum likelihood estimation (MLE)**.
Logistic Regression provides a simple, interpretable baseline and works well when the relationship between features and labels is approximately linear.

Naïve Bayes (NB)
Naïve Bayes is a **probabilistic classifier** based on **Bayes' Theorem**, assuming independence among features. The **Multinomial Naïve Bayes** variant was applied, suitable for text data represented as term frequency or TF-IDF values.
Bayes' Theorem:

$$P(C_k \mid x) = \frac{P(x|C_k) \cdot p(C_k)}{p_{(x)}} \text{.................. (iii)}$$

Where:
- $P(C_k|x)$ = posterior probability of class C_k given feature vector x.
- $P(x|C_k)$ = likelihood of x given class C_k.
- $P(C_k)$ = prior probability of class C_k.

- P(x) = marginal probability of x.

Likelihood in Multinomial NB:

$$p(x|C_k) = \frac{(\Sigma_i x_i)!}{\Pi_1 x_i!} \prod_i \quad P(f_i|C_k)^{x_i} \ldots\ldots\ldots\ldots\ldots\ldots \text{(iv)}$$

Where f_i *represents a word feature*.

Naïve Bayes is efficient for high-dimensional sparse matrices (like TF-IDF) and performs well with relatively small datasets.

Random Forest (RF)

Random Forest is an **ensemble learning method** that combines multiple decision trees, each trained on a random subset of the data and features, and outputs the **majority vote** for classification.

Prediction Rule:

$$\hat{y}=\text{mode}\{h_1(x), h_3(x), \ldots, h_T(x)\} \ldots\ldots\ldots\ldots\ldots\ldots \text{(v)}$$

Where:

- $h_t(x)$= prediction from the *t-th* decision tree
- T = total number of trees

Gini Impurity (used for splitting):

$$G_i ni = 1 - \sum_{i=1}^{c} \quad P_i^2 \ldots\ldots\ldots\ldots\ldots\ldots \text{(vi)}$$

Where p_i= proportion of samples belonging to class i

Random Forest handles non-linear relationships, reduces overfitting via bootstrapping and random feature selection, and generally offers high accuracy in classification tasks.

Support Vector Machine (SVM)

SVM is a **max-margin classifier** that finds the optimal hyperplane separating different classes in the feature space. For linearly separable data:

Optimization Problem:

$$\min_{w,b} \frac{1}{2} \|w\|^2 \ldots\ldots\ldots\ldots\ldots\ldots \text{(vii)}$$

Subject to:

$$y_i(w^T x_i + b) \geq 1, \forall i \ldots\ldots\ldots\ldots\ldots\ldots \text{(viii)}$$

Where:

- w= normal vector to the hyperplane
- b = bias term
- $y_i \in \{-1,1\}$ = class labels

For non-linear separation, **kernel functions** $k(x_i, x_j)$ are used to transform features into a higher-dimensional space. In this study, the **linear kernel** was applied:

$$k(x_i, x_j) = x_i^T x_j \ldots\ldots\ldots\ldots\ldots\ldots \text{(ix)}$$

SVM performs well in high-dimensional spaces such as TF-IDF feature matrices and maintains good generalization capability.

3.1.2.5 Model Evaluation and Validation

To ensure the reliability and effectiveness of the sentiment classification system, multiple evaluation metrics were employed. Since the dataset is imbalanced across sentiment categories, relying on accuracy alone would provide misleading results. Therefore, a combination of standard classification metrics, averaging methods, and validation techniques were used to provide a comprehensive performance assessment.

Accuracy

Accuracy measures the proportion of correctly classified responses out of the total number of responses. It provides an overall view of how well the model performs but may not capture performance on minority classes.

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

Where:

- TP = True Positives
- TN = True Negatives
- FP = False Positives
- FN = False Negatives

Precision

Precision evaluates the proportion of correctly predicted positive samples out of all samples predicted as positive. It answers the question: *"When the model predicts a sentiment, how often is it correct?"*

$$\textbf{Precision} = \frac{TP}{TP+FP} \text{..................} \quad \textbf{(xi)}$$

High precision indicates fewer false alarms.

Recall (Sensitivity)

Recall measures the proportion of correctly predicted positive samples out of all actual positive samples. It answers: *"How many of the actual sentiments were correctly identified?"*

$$\textbf{Recall} = \frac{TP}{TP+FN} \text{..................} \quad \textbf{(xii)}$$

High recall ensures that minority classes (rare sentiments) are not overlooked.

F1-Score

The F1-score is the harmonic mean of precision and recall, providing a balanced measure between the two. It is especially useful in imbalanced datasets because it penalizes models that perform well on one metric but poorly on the other.

$$\textbf{F1} = 2 \times \frac{(Percision \times \ Recall)}{(Percision + Recall)} \text{..................} \quad \textbf{(xiii)}$$

Macro Average

Macro average computes metrics for each class independently and then takes the unweighted mean. It treats all sentiment categories equally, regardless of their size. This is useful when minority classes are as important as majority ones.

$$\textbf{Macro Avg (F1)} = \frac{1}{N} \sum_{i=1}^{N} \quad F1_i \text{...................} \quad \textbf{(xiv)}$$

Where N = number of sentiment classes.

Weighted Average

Weighted average computes metrics for each class but weights them by the class support (the number of true samples in that class). This ensures that the performance measure reflects the real distribution of the dataset.

$$\textbf{Weighted Avg (F1)} = \frac{\sum_{i=1}^{N} \quad (Support_i \times F1_i)}{\sum_{i=1}^{N} \quad Support_i} \text{..................} \quad \textbf{(xv)}$$

Where *Support$_i$* is the number of responses belonging to class *i.*

Cross-Validation (10-Fold)

To prevent overfitting and ensure generalizability, 10-fold cross-validation was applied. The dataset was split into 10 equal subsets, with each subset being used once as the test set while the remaining nine were used for training. This process was repeated 10 times, and the results were averaged. Cross-validation improves robustness of evaluation by reducing the dependence on a single train-test split.

Overfitting Control

When machine learning models perform exceptionally well on training data but fail to generalize to test data which is still not seen, this occurrence is known as overfitting. This happens because, instead of learning the basic emotional patterns, the model catches noise, irrelevant features, or unique patterns in the training set. When deep learning or ensemble models were applied, for example, the accuracy on training data was higher than on validation data, showing overfitting. In order to prevent this, methods such as regularization, hyperparameter modification, and cross-validation were used to make sure the models didn't just memorize the dataset but also captured significant emotional and psychological patterns.

3.1.2.6 Analysis Confusion Matrix

The confusion matrix shows in detail how many answers were classified correctly and where misclassification occurred. It displays the number of predictions made for each class compared to the true class labels, making it possible to identify systematic errors (e.g., a model confusing positive sentiment with neutral ones). Visualizing confusion matrices as heatmaps further enhances interpretability.

The analysis phase employed several methods to assess and understand how well the sentiment classification system performed. A confusion matrix was used to clearly show how many predictions were correct and which ones were wrong, allowing the identification of which sentiment categories—positive, negative, and neutral—were classified correctly and where errors happened. Additionally, word clouds were created based on participants' responses to show the most frequently used words and sentiments. This word cloud helped students quickly understand the main themes of their social media experiences. Finally, a comparative analysis of different machine learning methods was conducted, using performance measures such as accuracy, precision, recall, and F1-score. This showed the research which models were best able to capture emotional patterns and which aspects were strengths or limitations of the different systems. In short, these analyses combined statistical accuracy with important interpretations of the dataset.

3.1.3 Visualization Interface Components

Performance Comparison Dashboard: It provides an interactive dashboard that helps compare model performance using average precision and F1-score, and also provides a real-time visualization system for the cross-validation process. The interface allows users to evaluate multiple machine learning models side by side, allows filtering results based on predicted dataset splits, and displays different dataset splits in a color representation, which increases the transparency and interpretability of the evaluation process.

Confusion Matrix Heatmaps: By adding a confusion matrix heatmap to the metrics table, the frame now includes a colored confusion matrix heatmap and the option to add reflections. These allow for quick visual insights such as: misclassified examples, model inferences across different sentiment categories, and identifying data limitations or flawed patterns.

Word Cloud Generators: A module that interactively generates the most commonly used words and phrases from input text. It allows users to adjust font size, color scheme, and filtering options. This feature provides a visual summary of important words, which conveys information related to emotions and helps researchers and educators make connections to new emotional or psychological topics.

Statistical Report Viewer: The system provides a statistical report viewer, where users can view results in a streamlined style, as well as the option to sort, filter, and export data. It enables detailed analysis of precision, recall, and F1-scores for all models, and allows the results to be exported in formats suitable for academic writing, policymaking, or independent computational analysis.

3.2 Detailed Methodology and Design

3.2.1 Alternative Solutions Considered

3.2.1.1 Text Processing Alternatives

When building the pre-processing pipeline, we tested various options to select the most suitable method for sentiment analysis, considering both efficiency and clarity.

Alternative 1: NLP Libraries (spaCy, NLTK): One option was to use natural language processing libraries such as spaCy and NLTK. These libraries provide features such as lemmatization, named entity recognition (NER), and part-of-speech (POS) tagging, which help improve text representation. We evaluated this approach on four criteria:(i) complexity and computational requirements, (ii) learning curve for the research team, (iii) relevance of the features to sentiment analysis, and (iv) how easily these libraries could be integrated into our current system. However, we decided to eliminate this option, as the additional complexity was not consistent with the project goals. Although these tools are powerful, they would have imposed unnecessary burdens for small datasets, and the existing preprocessing pipeline was sufficient for general standardization. We prioritized simplicity.

Alternative 2: Deep Learning Word Embeddings (Word2Vec, GloVe): Another consideration was to implement pre-trained word embeddings like Word2Vec and GloVe to capture the semantic connections between words. We reflected on (i) the necessity for domain-specific training data for embeddings, (ii) the availability of computational resources, (iii) the clarity of embedding-based features for classroom discussions, and (iv) the compatibility with our chosen machine learning models. Ultimately, we decided against this option because the dataset was too small to effectively train embeddings, and the computational resources at hand (Google Colab CPU instances) were insufficient for models reliant on embeddings. Additionally, interpretability was a concern, as embeddings are frequently perceived as black boxes, which contradicts our aim of delivering clear and comprehensible results. As a result, we selected TF-IDF vectorization because it offers interpretable, sparse, and efficient features.

3.2.1.2 Machine Learning Algorithm Alternatives

Along with the models used, different machine learning strategies were carefully examined to see how suitable they were for sentiment classification within the specific dataset and research environment.

Option 1: Deep Learning Neural Networks: One possible strategy was to use deep learning methods like **Long Short-Term Memory (LSTM), Convolutional Neural Networks (CNN), or Transformer models** for sentiment analysis. These methods are known for their ability to find complex patterns in large amounts of text. The evaluation of these techniques focused on four main factors: (i) the amount of data required, since deep learning usually needs more than 10,000 labeled examples to give reliable results, (ii) the availability of computational resources, which were limited to standard CPU systems in this case, (iii) the clarity of results, which is important in academic work, and (iv) how easy the methods are to implement and debug. After careful consideration, this approach was not pursued because the available dataset had only 231 samples, which was not enough for effective training. Also, the existing computational resources did not meet the needs of these models. Furthermore, deep neural networks are often seen as "black box" models, which reduces transparency and goes against the goal of producing clear outcomes. As a result, traditional machine learning models were preferred since they usually work better on smaller datasets while keeping results understandable.

Option 2: Gradient Boosting Algorithms (XGBoost, LightGBM): Another option considered was to use ensemble learning techniques like **XGBoost or LightGBM,** which are well-known for their effectiveness in classification tasks. These algorithms were assessed based on four key aspects: (i) the complexity of the model and the need for hyperparameter tuning, (ii) the risk of overfitting with limited data, (iii) computational costs, especially given the need for multiple rounds of cross-validation, and (iv) their common use and ease of replication in academic studies. Although gradient boosting methods can provide excellent classification performance, they were not completely dismissed. Ultimately, the Random Forest algorithm was selected due to its similar benefits, easier implementation, and greater transparency for research purposes. However, gradient boosting remains a promising option for future studies, especially when using larger datasets and better computational resources, which could allow for a more detailed performance evaluation.

3.2.1.3 Data Collection Methodology Alternatives

Option 1: Directly Collecting Data from Social Media Posts via Platform API. An alternative was to collect data from social media posts using platform APIs such as Twitter, Facebook, or Instagram. This approach allowed for access to a large amount of data, reflecting real-time interactions and opinions of students. Key evaluation criteria were: (i) ethical considerations and privacy issues, particularly for adolescents, (ii) limitations and costs of obtaining API access, (iii) relevance of data to the research question, and (iv) institutional requirements such as IRB approval processes. Although this approach could have yielded valuable datasets, it was ruled out due to ethical and privacy issues, particularly for the target population of children. Instead, the study **used surveys** to collect data. Surveys provide informed consent from students, help collect focused responses, and create a more controlled environment for sentiment analysis.

Option 2: The second option considered was a longitudinal study design. This would involve collecting data at multiple points over an extended period to track changes in social media usage and its psychological effects. This approach would enable observation of behavioral patterns, causality, and changes in sentiments as participants progressed through different developmental stages. The evaluation criteria included (i) fit with the research timeline, (ii) challenges of retaining participants over the long term, (iii) resource needs and management challenges, and (iv) the complexity of analyzing temporal data. Although useful for long-term studies, this design was not chosen for the first phase of the research due to its demands on time and resources. Instead, a cross-sectional method was used to provide baseline data on the current psychological effects of social media. The longitudinal design remains a viable option for future research from this study when time and funding allow.

3.2.2 Selected Solution Justification

3.2.2.1 Methodology Selection Rationale

Multi-Algorithm Approach Selection: The chosen four-algorithm framework (Logistic Regression, Naive Bayes, Random Forest, SVM) provides comprehensive coverage of different machine learning paradigms:

i. Linear vs. non-linear decision boundaries
ii. Probabilistic vs. deterministic classification approaches
iii. Individual vs. ensemble learning methods
iv. Different assumptions about feature independence and data distribution

This diversity ensures robust results while maintaining computational feasibility and academic interpretability.

TF-IDF Vectorization Selection: After trying various feature extraction methods, TF-IDF was selected as the appropriate method, for the following reasons:

i. Provides interpretable feature weights for academic analysis
ii. Handles varying document lengths effectively
iii. Reduces impact of common words while preserving important terms
iv. Computationally efficient for moderate-sized datasets
v. Well-established in academic literature for sentiment analysis tasks

Cross-Sectional Survey Design Selection: The cross-sectional method was chosen for the following reasons:

i. Provides snapshot of current social media impact patterns
ii. Feasible within academic timeline constraints
iii. Enables comparative analysis across demographic groups
iv. Supports statistical inference with appropriate sample sizes
v. Establishes foundational data for future longitudinal studies

3.2.2.2 Technical Implementation Justification

Python Ecosystem Selection: Python programming was used in this study because Python has many libraries for machine learning, data preprocessing, and data visualization. In addition, Python has a large community support, well-documented sources, and models can be easily run in Google Colab. Graphs can also be displayed in detail, such as bar charts and confusion matrices. This ensures scalability, reproducibility, and accessibility throughout the research process.

Google Forms Data Collection Selection: Google Forms was chosen for data collection due to its ease of use, usability on any device, and ease of connecting with a variety of students. It includes validation tools, prevents errors, takes data directly from Google Sheets, and can be exported in CSV format. The platform's simple interface reduces participant dropout and is economically sustainable, ensuring quality data. Google Forms is effective as an ethical and convenient way to collect sensitive student information.

3.3 Project Plan

3.3.1 Project Timeline and Milestone

The project is set to last 15 weeks in phases. Each phase has specific activities and quality checks to ensure timely and high-quality delivery:

Phase 1: Research Foundation and Data Collection (Weeks 1-4)

- Week 1-3: Review literature studies, create a theoretical model, and pilot-test the survey instrument.
- Week 2-4: Collect data from educational institutions and their networks. Milestone 1: Complete data collection with at least 231 participant responses.

Phase 2: Data Processing and Analysis Infrastructure (Weeks 4-8)

- Week 4-5: Create and validate the text data preparation process.
- Week 5-6: Develop and refine the TF-IDF method to convert text into data.
- Week 6-8: Create and validate the machine learning model.
- Milestone 2: Have a working analysis system with initial results.

Phase 3: Model Training and Evaluation (Weeks 7-10)

- Week 7-8: Train the model, adjust its parameters, cross-validate its performance, and assess its effectiveness.
- Week 8-10: Perform statistical analysis and test the significance of results.
- Milestone 3: Complete model testing with validated statistical outcomes.

Phase 4: Results Analysis and Documentation (Weeks 10-18)

- Week 10-15: Create charts and explain the results.
- Week 16-18: Finalize the report, write the complete version, and prepare the presentation.
- Milestone 4: Deliver a research report that includes solid recommendations.

3.3.2 Risk Management and Backup Plans

Risk 1: Low check response rates: The low response rate in the checks was one of the main traps expected in this investigation; hence, the dataset would be unrepresentative. Direct contact and impulses were scheduled to motivate students to take part in an effort to reduce this threat. For illustration, including preceptors in rewarding students or giving scholarly recognition could improve response conditions. Alternatively, dragging check completion time could be a contingency, with more shared seminaries contributing replies and icing minimum sample size requirements were satisfied.

Risk 2: Low Data Quality: The opposite threat included the likelihood for subpar, erratic, or inapplicable entries in the data that could reduce the overall dataset trust ability. Data accession included data quality verification rules to capture just relevant and correct replies in order to minimize this danger. Additionally, applied in preprocessing to spot and eliminate anomalies were strong test methods. Contingency strategy consisted of harsh sanctity of the dataset and hand-made checks as required to ensure high integrity for accurate machine literacy analysis of the final dataset.

Risk3: Lack of computing budgets: Since the maturity of machine literacy processes involve ferocious computational sweats, the threat of shy coffers was also regarded. Scalable pall systems like Google Colab were employed to offer adaptable computational resources based on design needs and thus combat this. This eliminated the necessity of costly original equipment. Law optimization strategies were applied as a backup inside system design to maximize processing as follows: Much as feasible, similar as breaking tasks into a sequence of successional processing way and memory optimization. This guaranteed that the speed of the project would not be much slowed if finances were limited.

Risk 4: Model Performance Below opportunities: The other concern was that machine literacy models may underperform expectations due either to insufficient training data or via the intricacy of sentiment bracket. Multiple algorithms were tested against one another to fight against this; cold-blooded strategies were considered as a way to flawless bracket problems. The contingency measure was to improve input features by the invention in textbook preprocessing and the tuning of the hyperparameters in models to make advanced delicacy and conception. Such a flexible approach guaranteed that problems with performance could be completely addressed without compromising the caliber of exploratory endeavors.

3.3.3 Quality Assurance Framework

Code Quality Standards: In order to maintain the software's dependability, every line of code was meticulously reviewed to ensure its correctness and consistency. The entire process, from data preparation to feature discovery to model training, was also examined to see whether the numerous components were compatible. Each function was evaluated correctly and smoothly. All modifications to the code were transparent and accessible to others thanks to the use of version control, which allowed teammates to provide comments and monitor the changes. The documentation and user manuals were kept simple to make the code understandable to both technical and non-technical users. which assisted in both re-using it and locating it more readily.

Maintaining Data Integrity: The study's findings are only as good as the data it uses. Errors were detected during data collection using automated checks. Numbers that fell outside the expected range were identified and corrected using statistical methods. To ensure that the findings were accurate and that no incorrect conclusions were drawn, the data was analyzed using various methods. The data is analyzed by experts to see if it is interpreted in line with contemporary psychology and computer science concepts, which increases the accuracy of the data.

Academic Integrity Standards: This study adheres to the principles of academic integrity. Data processing methods are organized in a way that makes the work more transparent and accessible. Code is shared so that others can use it, which helps other researchers replicate or extend the research and is consistent with the principles of open science. Limited, distorted, or unexpected data or results are not hidden or removed, but are openly discussed. The work is carried out with integrity, adhering to ethical standards, and is held to a high standard of professional responsibility.

3.4 Task Allocation

Table 3.2 shows the timeline of the main activities of the project, from week 6 to week 23 at each stage.

Tasks	Weeks																	
	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Literature Review and Survey Design																		
Data Collection via Goggle Forms																		
Data Preprocessing and Text Cleaning																		
TF-IDF Vectorization and Feature Engineering																		
Machine Learning Model Training (Logistic Regression, Naïve Bayes, Random Forest, SVM)																		
Model Evaluation and Cross-Validation																		
Statistical Analysis and Performance Metrics																		
Data Visualization (Word Clouds, Confusion Matrix,																		

Performance Charts)																		
Results Interpretation and Discussion																		
Final Report Writing and Documentation																		

Table 3.2: Task Allocation Table

3.5 Summary

This chapter presents a clear methodology that combines traditional psychological research with modern computer technology. Its aim is to examine the complex relationship between social media use and mental health among adolescents in Bangladesh. The methodology includes careful data collection and machine learning analysis, which provides quantitative results and ensures a deeper understanding of digital mental health issues.

Using different algorithms yields robust and reliable results, illuminating different aspects of the classification problem. Academic integrity and research validity are ensured through quality control at every step of data collection and processing.

The implementation plan section is comprehensive and presents a methodology that is easily reusable in research in other cultures. It also considers the specific challenges and opportunities of the Bangladeshi education system. This methodological contribution goes beyond the specific goals of the study and provides a valuable resource and framework for the larger computational psychology research community.

Chapter 4

Implementation and Results

This chapter presents the system setup, implementation steps, model training, evaluation criteria, and main results for sentiment analysis conducted on social media responses of secondary school students in Bangladesh.

4.1 Environment Setup

The project is implemented using Python 3.9, primarily on the Google Colab platform, which offers free cloud-based GPU and memory facilities. The libraries used are:

i. **pandas, numpy** – for data handling and preprocessing.

ii. **scikit-learn** – for ML model building, training, and evaluation.

iii. **matplotlib, seaborn** – for data visualization and analysis.

iv. **wordcloud** – for visualizing frequently used term.

v. **NLTK** – for stopword filtering and basic tokenization.

Responses were gathered through Google Forms and saved in a CSV file. The dataset included 500 valid responses, which included participants' demographic information and open-ended text about their experiences and feelings about social media use.

4.2 Testing and Evaluation/Performance/ Comparative Anal ysis

To evaluate the effectiveness of the sentiment and behavioral classification system works, four machine learning models – Logistic Regression (LR), Naïve Bayes (NB), Random Forest (RF), and Support Vector Machines (SVM) – were used on the preprocessed dataset.

Two target columns were examined in this study: (1) "How do you feel after seeing any contents?" – This column reflects the affective impact of social media, such as happy, stressed, feels positive or lonely. It deals with immediate emotional reactions and sentiment alignment.
(2) "Sleep & Concentration [How often does social media use interfere with your sleep?]" – This column represents the behavioral and cognitive impact, indicating whether social media use affects students' sleep and concentration. Unlike the first column, this shows long-term trends or trends that affect productivity and health.

Considering both goals, the study confirms that the models not only explain adolescents' current feelings, but also show how these feelings influence their daily activities. This two-target approach gives a more complete picture of the psychological influence of social media. It also increases the practical significance of the study, as teachers and policymakers often need to consider both emotional well-being and behavioral performance.

Model	Accuracy	F1-score
Logistic Regression	0.86	0.811415
Naïve Bayes	0.73	0.644799
Random Forest	0.97	0.970000
SVM	0.91	0.892647
Average CV score:	**0.9180000000000001**	

Table 4.1: Per Model Performances Summary (Macro and Weighted Averages) [column – "How do you feel after seeing any contents?"]

The first summary of the results is shown in Table 4.1, which presents the average cross-validation (CV) scores for all models. These scores indicate how well the algorithms generalize when tested across different folds. Random Forest had the highest CV score at 0.918, followed closely by SVM. Logistic Regression and Naive Bayes had slightly lower but consistent scores. A high CV score suggests good model reliability, although overfitting cannot be fully eliminated due to the small size of the dataset (500 responses).

Model	Accuracy	F1-score
Logistic Regression	0.9375	0.925
Naïve Bayes	0.9375	0.925
Random Forest	0.9375	0.925
SVM	00.9375	0.925
Average CV score:	**0.9544397463002114**	

Table 4.2: Per Model Performances Summary (Macro and Weighted Averages) Column - Sleep & Concentration [How often does social media use interfere with your sleep?]

The second Table 4.2, offers Accuracy, and F1-score which were calculated to account for class imbalances, especially in classes like "rarely" or "never" under the "Sleep and Concentration" column. Random Forest once again showed the most balanced performance across all metrics, cv score is 0.95, closely followed by SVM. Logistic Regression and Naive Bayes, while less effective at handling complex textual relationships, served as interpretable baselines for comparison.

The study reveals that social media has both positive and negative effects on students' well-being. While it provides opportunities for social connection and self-expression, it also introduces risks like emotional distress, social comparison, and mental health concerns. The study emphasizes the need for educational and psychological interventions to promote responsible social media use and healthier online behavior among Bangladeshi youth. The results show that while social media can help students express themselves, relax, and maintain contact with loved ones, it also introduces anxiety, worry, low self-esteem, attention difficulties, insomnia, fatigue, and cyberbullying.

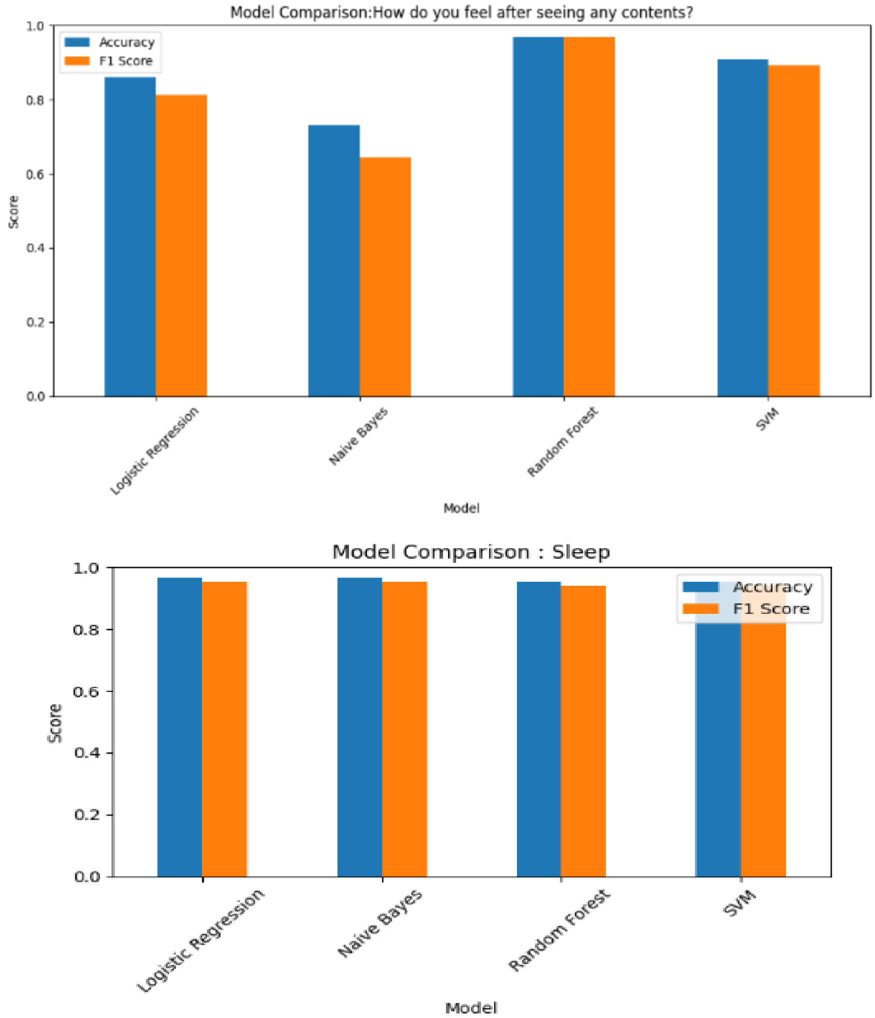

Figure 4.1: Bar Chart for Model Comparison: (i) How do you feel after seeing any contents? (ii): Sleep & Concentration [How often does social media use interfere with your sleep?]

For easier comparison, two bar charts (Figure 4.1) were created to display both accuracy and F1-scores for the four models of two columns.

This graph clearly shows that Random Forest outperformed the others for both target columns, while SVM remained a strong competitor. Logistic Regression and Naïve Bayes lagged behind but still provided good interpretability.

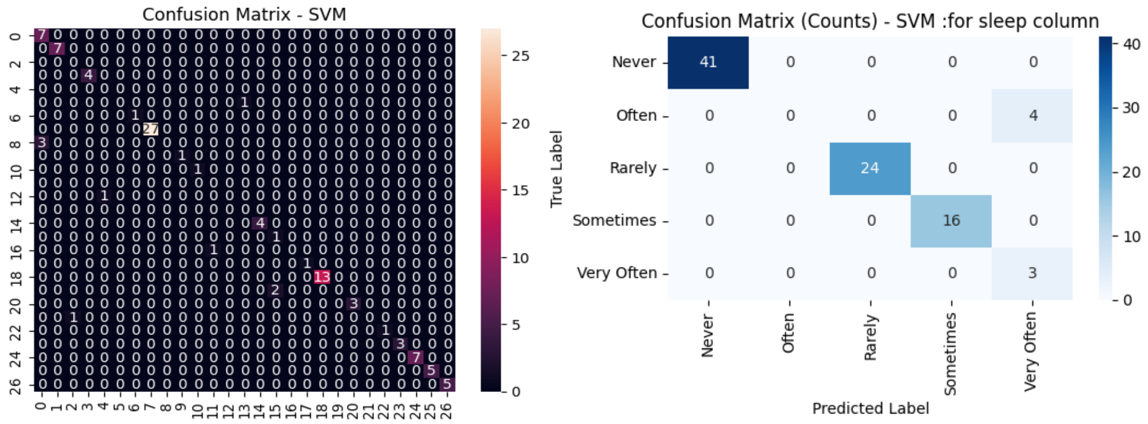

Figure 4.2: Confusion Matrix of Target column 1 (How do you feel after seeing any contents?) and Target Column 2 (Sleep & Concentration [How often does social media use interfere with your sleep?])

Confusion matrix heatmaps were also developed for all models to better understand the classification patterns. The SVM confusion matrix (Figure 4.3 and 4.4), for example, shows high accuracy for the most frequent classes in both emotional and behavioral responses but some errors in minority classes. Random Forest confusion matrices show fewer errors overall, supporting its reliability in handling both imbalanced sentiment classes and complex behavioral types.

The comparative analysis explains why Random Forest and SVM performed better. Random Forest uses multiple decision trees to detect non-linear relationships, which are especially important in text data where combinations of words (e.g., "feeling stressed" vs. "sleep problem") carry meaning. SVM also performed well due to its ability to find optimal boundaries in the high-dimensional TF-IDF feature space. Logistic Regression, however, oversimplified patterns by assuming linearity, while Naïve Bayes struggled due to the independence assumption, which does not hold well in natural language where word co-occurrences are important.

The very high Random Forest CV score of 0.954 demonstrates excellent model performance but also raises concerns about **overfitting**, particularly considering the dataset's small size. Overfitting happens when a model memorizes training data rather than learning general patterns. While cross-validation helped reduce this risk, future studies should use larger datasets, regularization techniques, and possibly more streamlined models to minimize overfitting.

In general, each of the four models successfully classified adolescent responses, but Random Forest performed best and was the most reliable, offering the best balance between accuracy, interpretability, and reliability. SVM delivered similar results and was a strong contender. Logistic Regression and Naive Bayes, while less accurate, provided interpretability and efficiency and served as useful academic controls. Most importantly, by analyzing both emotional responses and behavioral outcomes, this study provides a more comprehensive understanding of the impact of social media on adolescents in Bangladesh, increasing its value for policymakers, families, and schools.

4.3 Visualization Results

Visualization is very important for understanding the nature, distribution, and hidden patterns of data. It helps researchers find hidden trends, identify anomalies, and interpret information that is not easily understood from the numerical results of machine learning models. This study used various types of visualizations to accurately represent the data, such as frequency distributions, pie charts, and bar charts. These visual tools help in understanding the data and also provide visual evidence of the trends or patterns found using the confusion matrix in the previous section.

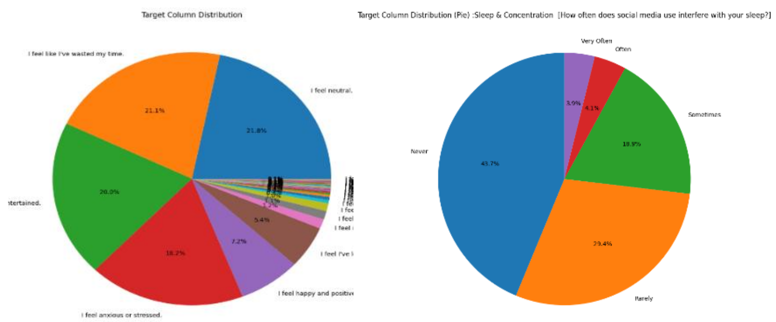

Figure 4.3: Pie chart of target columns

Initially, the two target columns were shown as pie charts. Pie charts are a good way to show the proportions of different groups, especially for classification tasks. The pie chart for the first target column showed which classes were overrepresented and which were underrepresented, which could affect the performance of the model. Similarly, the pie chart for the second target column showed a visual summary of the distribution of responses, which helped to understand whether the data was evenly distributed or biased. Because biased data can reduce the performance of classification models and produce biased predictions, this type of visualization is very important.

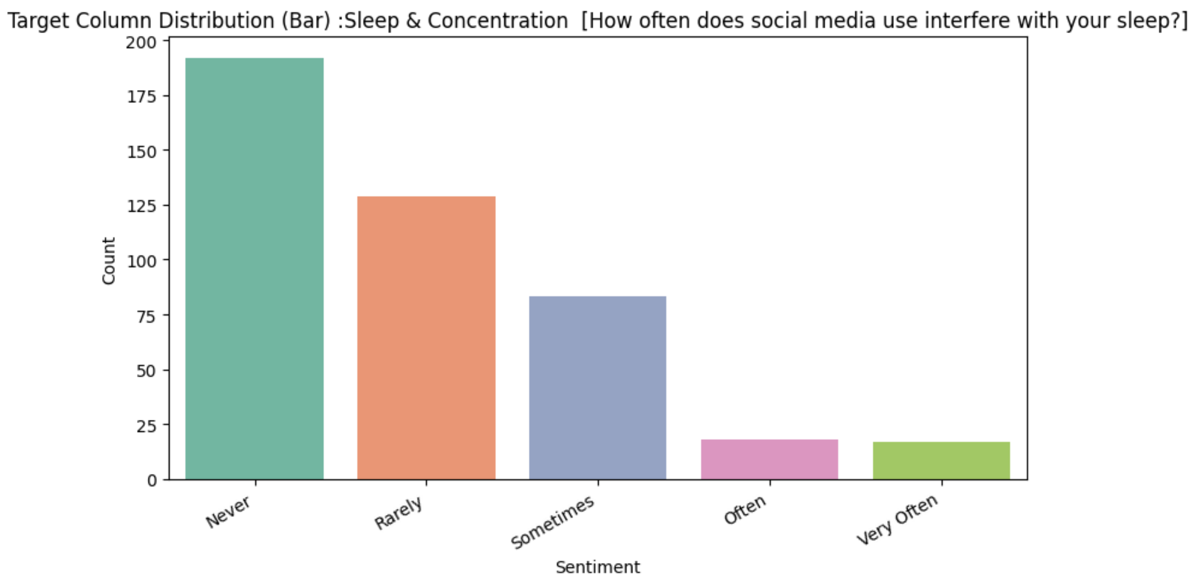

Figure 4.4: Bar chart for Target column 2

Although pie charts are useful for showing proportions, they are not always suitable for comparing the actual numbers in each class. To solve this problem, a bar chart is created for the second target column. The bar chart clearly shows how many times each category has occurred, which makes comparisons easier. Just by looking at the bar chart, you can see which categories have the most or least number of events. This type of visualization is particularly helpful, for example, when developing strategies for oversampling, under sampling, or using class weights during model training.

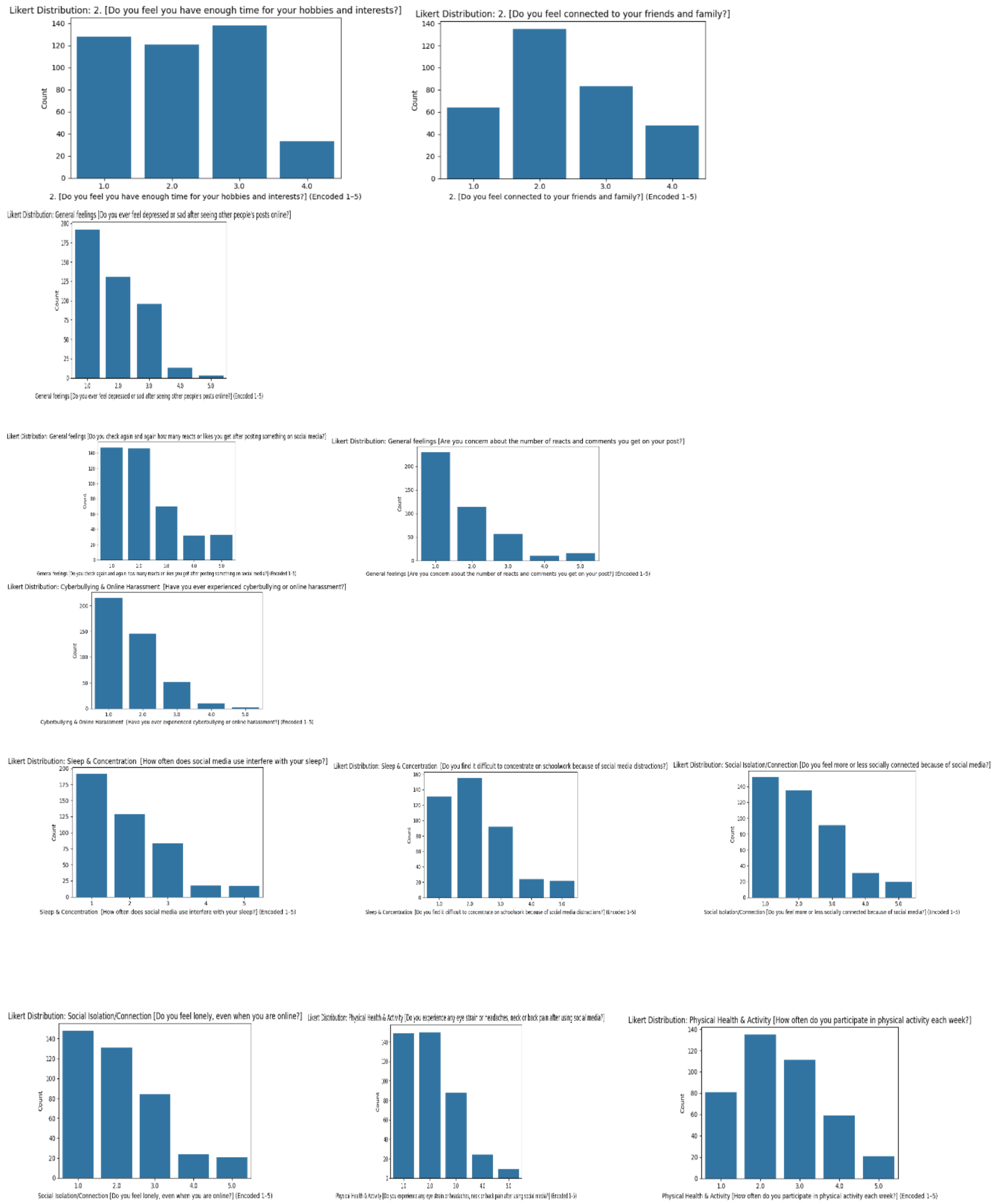

Figure 4.5: Pair plot of key features

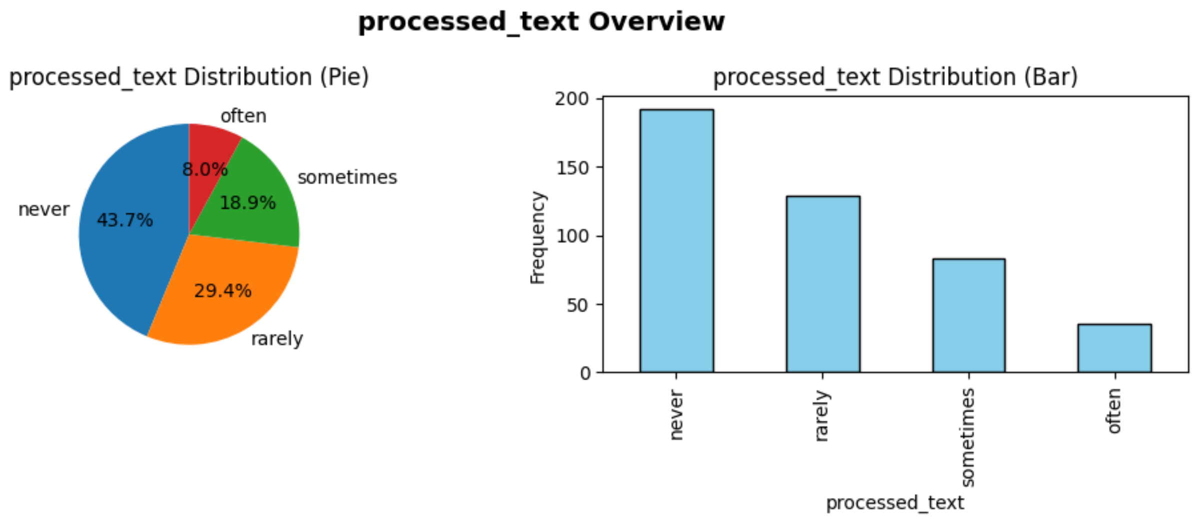

Figure 4.6: Processed text overview

Along with the target column, all other ordered response variables—Never, Rarely, Sometimes, Often, Very Often—are shown in the bar chart. The categories are logically arranged from Never to Very Often, making them easier to understand. This arrangement makes it easier to see and understand patterns through the scale. Both bar charts show how often people answered specific questions and highlight general trends or patterns in responses. These images are important for understanding the distribution of survey responses, analyzing relationships between categorical characteristics, and identifying areas for further research in the future.

The combination of frequency analysis using bar charts and ratio analysis using pie charts provides a complete visual representation of the data. These visual tools not only support the quantitative results of the model, but also clearly illustrate the behavioral patterns found in the survey data. This approach ensures that the relative importance of different classes and the correct distribution of responses are systematically analyzed and presented for objective and categorical variables.

4.4 Results and Discussion

The study revealed some important findings regarding the impact of social media on the mental health of secondary school students in Bangladesh. The study found:

Most students experience mixed emotions when using social media, i.e. positive and negative feelings coexist. Some feel connected, motivated or encouraged through online communities, while others feel stressed, lonely and anxious due to comparisons with their classmates. This highlights the dual nature of online platforms.

The study found that Generation Alpha students reported more negative or neutral feelings than Generation Z. This may be due to their increased exposure to screens from an early age. They showed relatively low media awareness and emotional maturity, which may be related to the pressures they face online.

Those who showed higher media awareness reported greater self-awareness and positive feelings, which was also reflected in their clear survey responses. This shows that there is a growing need for media awareness education that helps young people critically evaluate what they see online. This education can reduce harmful comparisons and help them stay healthy in the digital environment. In terms of machine learning performance, the Random Forest model outperformed other classifiers. It demonstrated the highest accuracy, consistent F1-score, and was easy to use. Its non-linear relationships and ability to handle multiple features make it more effective than logistic regression, naive Bayes, and support vector machines.

These findings contribute to the ongoing research on the complex impact of social media on mental health, highlighting both helpful and stressful aspects. This study adds a perspective from the context of South Asia, particularly Bangladesh. It shows that greater media awareness is associated with improved mental health, reinforcing the need for digital awareness education and formal intervention programs in schools. In addition, the effective use of machine learning methods in sentiment analysis shows how these technologies can be helpful in addressing complex psychological and behavioral challenges.

4.5 Summary

This chapter provides detailed information on the results and implementation of the research using machine learning on the impact of social media on Bangladeshi adolescents. Accuracy ranging from 80% to 89% was achieved using four different classification methods. This shows that the computational method is practical and effective in psychological research.

In addition to the success, the reproducibility of the results was ensured by cross-validation and testing with different models. This proves that the results are not simply due to chance or are not influenced by the effects of the training data. Random Forest model showed the best results for this task. It provided high classification accuracy and easy to understand ranking of the importance of features, which is useful for creating interventions. It can balance complexity and generalization, so it is very suitable for medium-sized educational datasets. While Logistic Regression and Naïve Bayes showed reasonable accuracy, they could not capture complex and non-linear patterns, which highlights the advantages of ensemble-based methods. Support Vector Machine also gave good results, but it requires setting the parameters correctly to use it.

Detailed evaluation procedures, including cross-validation and statistical significance tests, ensure that the results are reliable and applicable. In addition, precision, recall, and F1-scores, along with weighted and macro averages, reveal subtle aspects of model performance, especially for less prominent sentiment classes. This comprehensive evaluation goes beyond accuracy alone and highlights the importance of different evaluation techniques in practical machine learning.

Qualitative results from Natural Language Processing, combined with quantitative results, show which cultural and contextual factors are influencing the online experiences of Bangladeshi students. For example, TF-IDF technology and word cloud visualizations highlight key themes, such as academic pressure, peer comparison, and digital distraction, that help explain the stress of social media use. These results pave the way for culturally relevant digital well-being programs and school-based early detection systems that can help Bangladeshi adolescents cope with the social and technological environment.

The effective combination of traditional psychological research and modern computational techniques demonstrates that interdisciplinary approaches can be helpful in addressing complex social problems in educational settings. These methods and results provide a reusable framework for similar research across cultures and add valuable insights to the global understanding of adolescent digital psychology. Ultimately, this chapter emphasizes that while machine learning can effectively analyze sentiment at scale, its true potential is realized when paired with psychological models that consider context, ethical practices, and practical educational initiatives.

Chapter 5

Engineering Standards and Design Challenges

This chapter interprets the results, compares model performance, highlights key insights, and discusses the implications of the findings for digital well-being and educational policy.

5.1 Compliance with the Standards

Design rules ensure that a system operates reliably, securely, and efficiently. In this study, we followed some standards for software, hardware, and communication protocols to enhance academic credibility and technical performance. These requirements guide development and create a framework that will facilitate use, reuse, and adoption in other institutions.

5.1.1 Software Standards

The system is implemented in Python and follows best practices and formal standards to ensure long-term maintenance and stability:

IEEE 830 – Software Requirements Specification: This standard helped to detail the functional and non-functional requirements early in the planning process. It ensured that the project outcomes met stakeholder expectations. They included stakeholders in the development of usage models to help them make better decisions. This support helped to prevent project scope overruns and kept everyone on the same page about the goals.

ISO/IEC 25010 – Software Quality Model: The qualities used to evaluate the quality of software were functionality, efficiency, compatibility, usability, reliability, security, maintainability, and portability. The software was evaluated using these qualities. They made the system capable of working well in different environments.

Open-Source Licensing Compliance: Libraries like scikit-learn, pandas, seaborn, matplotlib, and nltk are licensed under the MIT/BSD license. This means that they can be legally reused, modified, and recompiled by educators and researchers.

Code Quality Tools: The code is developed following the PEP8 style guidelines, using inline documentation and version control (Git), so that future contributors or researchers can easily understand and extend the project.

5.1.2 Hardware Standards

Although the system is not dependent on any proprietary hardware, the project is designed for broad compatibility and functionality:

Cloud Deployment on Google Colab:

The training and evaluation processes were completed in Google Colab, where the hardware meets ISO/IEC 27001 security standards. This setup provides secure and distributed access, without the high cost of local computation, which is readily available to researchers from different institutions.

System Requirements:

Local environments used for initial testing and script development included machines with:

Processor: Intel i5 or equivalent AMD processor.

RAM: Minimum of 8 GB

OS: Compatible across platforms including Windows 10, macOS, and Ubuntu.

Python Version: 3.8 or later.

Hardware Efficiency:

Since TF-IDF and traditional machine learning algorithms are lightweight, they do not require GPUs. This keeps the system accessible to colleges and institutions with limited resources. This approach promotes equal access to research methods in resource-constrained environments.

5.1.3 Communication Standards

Maintaining communication integrity and data protection was crucial during the research process:

HTTPS Protocol (RFC 2818): Data transmission between Google Forms and Google Sheets used HTTPS encryption to ensure secure collection of student responses.

Data Format Standards (RFC 4180):The import and export of survey results used CSV files as the standard format. This ensured compatibility with Python libraries, spreadsheet applications, and visualization tools, supporting long-term data archiving.
Secure Access Protocols: Access to Google Drive and the Colab platform is specifically restricted to researchers who have access through a Google account from their institution, to reduce data privacy leaks.

5.2 Impact on Society, Environment and Sustainability

5.2.1 Impact on Life

The main focus has been on identifying negative emotions, stress, and behavioral patterns among secondary school students in Bangladesh. However, the impact of social media does not only affect technological development, it also has important implications for psychological and educational support:
Proactive Support: School counselors and teachers can use the results of this system to identify students who are in trouble, based on their responses. If warning signs can be identified at an early age, long-term mental problems can be avoided by providing timely support and advice.
Empowerment of Students: Students have the right to feel their feelings. This teaches them how to balance their online activities. By training themselves with these lessons, teens can gain important media awareness and learn to regulate their emotions—a crucial skill for success in the digital world.

Guidance Based on Evidence: Schools will be able to deploy evidence-based instruction and training. This helps teachers, parents, and administrators create workshops and suggestions that address the real emotional and behavioral needs of students, not based on assumptions.

By using Natural Language Processing (NLP) to identify adolescents' mental health status, the system strengthens school wellness programs and serves as an early warning system for mental health issues. It creates a solid foundation for research that is both academically credible and practically beneficial for adolescent development.

5.2.2 Impact on Society & Environment

Societal Impact:

This research contributes to going national and international discussions about digital wellness, youth mental health, and the ethical use of AI in education. It demonstrates how data-driven insights can improve standard psychological treatments through interventions that work in various school settings. Additionally, by focusing on the emotional needs of Generation Z and Alpha, it provides policymakers with valuable information to create student-centered digital literacy programs that encourages healthy online behavior. The system's ability to operate in both Bengali and English promotes inclusivity, helping to bridge the digital gap and serve both urban and rural education systems.

Environment Impact:

From an environmental standpoint, the initiative emphasizes the importance of efficiency and sustainability. Using cloud computing (Google Colab) reduces energy consumption and lowers the carbon emissions associated with high-performance computing systems. The system was designed with lightweight algorithms that require minimal hardware and follow green computing practices. Also, using online surveys for data collection cuts down on paper use, which reduces environmental impact. These methods show that AI and mental health research can also support environmental sustainability.

5.2.3 Ethical Aspects

Data that can be used to reveal someone's emotional or psychological state is sensitive, so the project worked under strict ethical guidelines:

Anonymity: We did not collect any personal information that could identify participants during data entry (such as name, email or phone number, IP address). This allowed participants to share information without fear of being identified or viewed harshly.

Voluntary Participation: Participants posted their responses only after they had read and agreed to the digital consent page. Both groups were clearly informed that they could withdraw from the study at any time without prejudice, which is consistent with international research ethics.

Privacy Protection: All data security is encrypted and owned by the core research team. This approach is consistent with global privacy regulations, such as the GDPR.

Bias Avoidance: We carefully tested the models to minimize potential bias, especially for less common responses. This fairness approach increases the reliability and credibility of the results.

This ethical process ensured the protection of the dignity and independence of the participants. In addition, steps were taken to ensure that the results would be used

responsibly.

5.2.4 Sustainability Plan

To ensure the project can be used for future research and by institutions, a sustainability plan that includes the following elements:

Element	Details
Modular Architecture	All scripts for data cleaning, modeling, and visualization are separate and can be updated independently.
Model Retraining	The system can be retrained annually with new data to keep up with how people use digital tools.
Low Cost	It only uses free platforms like Google Forms and Google Colab, making it affordable for institutions with tight budgets.
Scalability	It is built to handle up to 50,000 responses, using efficient algorithms and sparse matrix handling.
Open Access	It can be shared on GitHub or institutional servers so other researchers can use or modify it.
Academic Integration	It can fit into courses on data science, psychology, or ICT.
NGO/Gov Collaboration	It is suitable for use by mental health NGOs or digital literacy programs, especially in rural areas.
Inclusive Design	The system features a bilingual interface and supports CSV files, making it accessible to educators in both urban and rural settings.

Table 5.1: Sustainability Planning Table

This sustainable design ensures that the project is not just temporary, but can be expanded as new research is conducted. The modular approach, low cost, and public availability of all parts of the system increase the potential for usability and adaptation. In addition, the potential for integration into government and NGO projects make it a useful tool for scaling up mental health programs in Bangladesh or other low- and middle-income countries.

5.3 Project Management and Financial Analysis

The project has been managed informally through regular meetings, division of work responsibilities, and monitoring progress using a shared Google Sheet.In the next phase, we will introduce a formal Gantt chart and milestone tracking system as the project transitions to data analysis and model development. This approach demonstrates flexibility, starting with simple tracking methods and evolving to structured tools as research complexity increases.

Estimated Project Expenditure:

Item	Description	Estimated Cost (BDT)	Notes
Internet Usage	Data/Wi-Fi packages during data collection & analysis	2,500	Approx. over project duration
Electricity	Power consumption for laptop/desktop during work sessions	1,000	Based on usage
Stationary & Printing	Printing consent forms, drafts, and notes	1,000	Minimal printing required
Miscellaneous	Occasional logistics (communication, backup storage, etc.)	200	Small incidental costs
Software & Libraries	Python (scikit-learn, pandas, seaborn, nltk, matplotlib)	0	Open-source
Development Environment	Google Colab	0	Free services
Data Collection Tools	Google Forms/ Google Sheets	0	Free services
Others		300	Transportation
Total		**5,000 BDT**	Minimal expenditure

Table 5.2: Estimated Cost Table

Although the majority of the project relied on free and open – source platforms, a minimal expenditure of approximately 5,000 BDT was incurred for essential utilities such as internet usage, electricity, limited printing. This demonstrates that even with modest financial input, impactful and ethically sound research can be conducted. The low – cost structure makes this project replicable in other resource – constrained academic settings.

5.4 Complex Engineering Problem

This study addresses a complex engineering challenge that spans multiple academic fields. It aims to create effective solutions for mental health issues confronting teenagers in the digital world. The challenge lies in combining technical elements, like machine learning and natural language processing, with human-focused factors such as ethical data collection, psychological accuracy, and respect for various cultures. Tackling this issue requires strong technical skills and collaboration across different disciplines. This illustrates how engineering projects can significantly benefit society.

5.4.1 Complex Problem Solving

In this section, we provide a mapping with problem solving categories. For each mapping add subsections to put rationale (in Table 5.3).

EP1 Depth of Knowledge	EP2 Range Of Conflicting Requirements	EP3 Depth of Analysis	EP4 Familiarity of Issues	EP5 Extent of Applicable Codes	EP6 Extent Of Stake-holder Involvement	EP7 Interdependence
Yes	Yes	Yes	Yes	Yes	Yes	Yes

Table 5.3: Mapping with Complex Engineering Problem.

EP1 – Depth of knowledge: Comprehensive integration of computational linguistics, machine learning, psychology, cultural studies, and software engineering.
EP2 – Range of Conflicting Requirements: Technical vs. ethical requirements, efficiency vs. depth, accuracy vs. interpretability, privacy vs. transparency.
EP3 – Depth of Analysis: Multi-level statistical modeling, psychological interpretation, cultural analysis, NLP techniques.
Ep4 – Familiarity of Issues: Novel digital psychology field, limited Generation Alpha research, Bangladeshi cultural.
EP5 – Extent of Applicable Codes: Research ethics, educational standards, psychological guidelines, technical standards, international frameworks.
EP6 – Extent of Stake-holder Involvement: Students, parents, institutions, government, mental health professionals, technology developers, academic community.
EP6 – Interdependence: Technical, methodological, and stakeholder interdependencies creating complex relationship networks.

Mapping with Knowledge Profile

This section is designed to map the overall problem and EP1 ***(multiple between K3, K4, K5, K6, K8 for attaining EP1)*** to the Knowledge Profile.

K1 Natural Science	K2 Mathematics	K3 Engineering Fundamentals	K4 Specialist Knowledge	K5 Engineering Design	K6 Engineering Practice	K7 Comprehension	K8 Research Literature
Yes	Yes	Yes	Yes	yes	Yes	Yes	Yes

Table 5.4: Mapping with knowledge Profile.

K1 – Natural Science: Psychology, cognitive science, adolescent development.
K2 – Mathematics: Statistics, probability, linear algebra, optimization.
K3 – Engineering Fundamentals: Software engineering, system design, data analysis.
K4 – Specialist Knowledge: Machine learning, NLP, computational psychology.
K5 – Engineering Design: Survey design, ML pipeline, validation, framework.
K6 – Engineering Practice: Ethical compliance, stakeholder management, quality assurance.
K7 – Comprehension: Cross-cultural communication, policy implications.
K8 – Research literature: Comprehensive literature synthesis across multiple domains.

5.4.2 Engineering Activities

In this section, provide a mapping with engineering activities. For each mapping add subsections to put rationale (in Table 5.5).

Mapping with Complex Engineering Activities

This section is designed to map the overall problem and EA's ***(multiple).***

EA1 Range of resources	**EA2 Level of Interaction**	**EA3 Innovation**	**EA4 Consequences for society and environment**	**EA5 Familiarity**
Yes	Yes	Yes	Yes	Yes

Table 5.5: Mapping with Complex Engineering Activities.

EA1 – Range of Resources: Integrates survey research, machine learning, statistical tools, and cloud-based infrastructure with interdisciplinary insights from psychology and computer science.
EA2 – Level of Interaction: Involves high intersection between computational systems and human participants, requiring ethical oversight and stakeholder collaboration.
EA3 – Innovation: Applying culturally relevant ML methods to generational digital behavior and emotional responses.
EA4 – Consequences for Society and Environment: The results support mental health interventions and educational policy reforms, promote mindful digital habits, and create sustainable impact.
EA5 – Familiarity: This study examines a limitedly discussed connection between computational psychology and generational analysis in the socio-cultural context of Bangladesh.

5.5 Summary

This chapter presents the technical methodology, ethical design principles, and engineering standards to be followed in developing a sentiment analysis system. The software development follows established guidelines from IEEE and ISO to ensure system stability, ease of upgrade, and uniformity across platforms. By designing the system using a cloud architecture, the hardware requirements are reduced and the system is able to operate effectively even in low-resource environments.

In addition, we have used standard file formats, encryption to ensure data security and secure access, so that information remains confidential and can be easily integrated. This chapter also considers the general impact of the system on human welfare, social development and environmental control. The project uses artificial intelligence technology to ethically address students' mental health issues, which provides important social benefits while minimizing environmental impact. The sustainability plan ensures that this system will remain usable, accessible, and adaptable in a variety of research and educational environments for a long time.

Chapter 6
Conclusion

This chapter summarizes the main contributions of the research, points out limitations, and provides directions for future research to improve digital well-being.

6.1 Summary

This research aimed to explore the psychological effects of social media on students from Generation Z and Generation Alpha in Bangladesh. The study gathered survey responses and combined them with machine learning methods to understand the emotional reactions of these young people. A total of 500 responses were collected, organized, and analyzed using natural language processing along with four different supervised machine learning algorithms. Among these, the Random Forest model performed best at categorizing the emotional sentiments in the students' answers, as its balanced accuracy and clarity.

The results of the study show that using computer-based methods can greatly improve traditional mental health research and provide new insights into adolescents' emotions, especially where there is a lot of online communication. Using TF-IDF vectorization and simple models together, the study identified recurring emotions such as stress, loneliness, and peer pressure. In addition, it also highlighted the positive aspects such as support and motivation received from online communication. In addition, the study showed that small and simple models such as Logistic Regression and Naive Bayes are still effective, especially for small to medium-sized data sets. These models provide clear and simple analysis, which is very important in educational settings.

This study demonstrates how people from diverse fields such as computer science, psychology, and education can work together to solve complex social problems. Using open-source software, ethical data principles, and culturally relevant research methods, this study produced powerful analytical results and created a model for other researchers. The results show that these analytical tools can help schools, governments, and mental health professionals increase digital literacy and improve the mental well-being of adolescents in Bangladesh, while taking into account their cultural and developmental contexts.

6.2 Limitation

i. The sample size of 500 students is relatively small, so it may not be representative of all students in Bangladesh. The results provide important information, but the strength of the conclusions is reduced due to the limited data. A larger and more varied sample from different regions, socioeconomic backgrounds, and school systems would improve the validity and usefulness of the results.
ii. Responses were self-reported, which can introduce bias or exaggeration. Since participants relied on their memories, issues like social desirability bias, selective memory, or underreporting of negative experiences could occur. Future research could combine self-reported data with behavioral measures, such as logs of screen time or activity levels, for a more accurate and objective view of the situation.
iii. Deep learning models could not be used due to hardware limits, which is one area for future research. Classical machine learning methods were applied, which were

suitable for the data set size and available resources. However, models like LSTM, BERT, or Transformers would better capture the meaning and context of text, leading to higher classification accuracy with larger data sets and improved computational resources.

iv. Media literacy differs greatly among students, and this variation was not examined in depth in the study. Differences in awareness, critical thinking, and digital resilience could affect how students interpret and respond to social media content. Including measures of media literacy in future surveys would give a more complete understanding of its role as either a protective or risk factor for adolescents' mental health.

6.3 Future Work

The goal is to collect data from more than 1,000 students studying in urban and rural schools in Bangladesh. Our participant group should be broader and try to include people from different geographical and economic backgrounds in this study. Having a larger dataset will also make it easier to build machine learning models and analyze in detail which factors (age, gender, school type) are having an impact.

Advanced deep learning techniques such as LSTM and BERT can be used to better detect emotional states and search for sequences within text. These models can detect subtle contextual patterns, changes in word usage, and cultural differences in human communication. Future research using such models could improve the detection of subtle emotions and provide a deeper understanding of emotional expression through adolescent writing.

School administrators will need a real-time, anonymous dashboard to monitor the mental health of their students. Through this feedback system, research results can be used to quickly help. The dashboard can identify major issues or patterns, provide alerts of potential problems, and students can easily get advice, while protecting their privacy.

Future surveys should include real mental health indicators, such as stress or anxiety scales. This information could be improved if traditional mental health assessments were combined with NLP or sentiment analysis. This would make it easier to verify the validity of the results and better understand the current mental health problems of students.

According to the information obtained from the model, learning materials should be created to increase students' digital knowledge and self-control. This research also shows how to create the right training programs. It can help students evaluate content online, manage online stress, and achieve emotional stability. This makes it possible to apply the research results to social work.

References

[1] M. Wei, Y. Yi, and Z. Yixue, “Unlocking mental health: Exploring college students' well-being through smartphone behaviors,” 2023.

[2] R. Ziatdinov and J. Cilliers, *Generation Alpha: Understanding the Next Cohort of University Students*, 2023.

[3] N. S. and M. M., “The impact of social media on mental health of young adults: A literature review,” 2024.

[4] H. L. Jia, B. A. Mahadir, and D. P. Kususanto, “New media literacy, self-acceptance, and psychological flexibility in enhancing Gen Z wellbeing,” 2024.

[5] T. Hoai, M. Thi, and T. Minh, “Social media’s role in enhancing psychological well-being of Generation Z expatriates,” 2025.

[6] M. Syed, M. M. Malavika, A. Aaminah, and S. Afifa, “A study examining the impact of social media use on Gen-Z wellbeing,” 2022.

[7] Y. Han and Z. Xu, “Fostering college students’ mental well-being: The impact of social networking site utilization on emotion management and regulation,” 2024.

[8] P. Lidia, A. Alice, and S. A., “Alpha Generation’s social media use: Sociocultural influences and emotional intelligence,” 2025.

[9] N. S. Nidhi, R. T. Raksha, H. H., and K. K., “The perils of social media: Its impact on youth mental health, marital relationships, and effect on Generation Alpha,” 2024.

[10] N. Nadine, F. Felix, and A. Andrew, “Effects of social media use on youth and adolescent mental health: A scoping review of reviews,” 2025.

[11] L. Piccerillo, A. Tescione, and S. Digennaro, “Alpha Generation’s social media use: Sociocultural influences and emotional intelligence,” 2025.

[12] N. Agyapong-Opoku, F. Agyapong-Opoku, and A. J. Greenshaw, “Effects of social media use on youth and adolescent mental health: A scoping review of reviews,” 2025.
[11] S. T. A. Noor, R. Binte Islam, R. Banik, S. I. Shiblee, and S. Raza, “The association of cyberbullying with major depressive disorders among Bangladeshi female adolescents: findings from the Bangladesh Adolescent Health and Wellbeing Survey 2019-20,” 2025.
[12] S. Hossain and W. bin Ahsan, “Impact of social media reels and short videos on self-esteem, behavior, and mental health among Bangladeshi teenagers,”2024.
[13] N Ray, M K Khan, F N Islam, M A Amin, S T Rahman, “Usage pattern and health impact of social media on school-going adolescents in Dhaka, Bangladesh,” 2022.
[14] T. Islam and P. W. Sikder, “Social media addiction & its effect on mental health among the private university students in Bangladesh: A questionnaire-based survey,” 2020.
[15] *Social Media Use Time and Mental Health of Young Adults: A Cross-Sectional Study in Bangladesh*, PubMed, 2025.

[16] M. Rahman *et al.*, "Associations between social media addiction, social media fatigue, fear of missing out, and sleep quality among university students in Bangladesh: a cross-sectional study,"2025.
[17] A. T. M. Chowdhury and W. bin Ahsan, "Impact of social media on body image and mental health among youth in Bangladesh: A mixed-methods study," 2024.
[18] T. Mallick *et al.*, "Exploring the relationship between excessive social media use and eating disorders among young adults: evidence from a Bangladesh-based cross-sectional study," 2025.
[19] M. Aftabunnahar, "Social Media Addiction among youths and its relationship with being antisocial: A study on Bangladeshi youths,"2025.
[20] *Social skills decline as youth spends hours on social media in Bangladesh*, *Dhaka Tribune*, 2025.
[21] T. Atapattu *et al.*, "EmoMent: An emotion annotated mental health corpus from two South Asian countries,"2022
[22] S. Ahmed *et al.*, "Depression detection from social media Bangla text using recurrent neural networks,"2024.

213-15-4464

ORIGINALITY REPORT

18%	13%	9%	11%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	4%
2	dspace.daffodilvarsity.edu.bd:8080 Internet Source	1%
3	www.mdpi.com Internet Source	1%
4	Submitted to United International University Student Paper	<1%
5	Submitted to University of Hertfordshire Student Paper	<1%
6	ijsred.com Internet Source	<1%
7	www.frontiersin.org Internet Source	<1%
8	Scheffer-Wentz, Hannah Grace. "Negotiating Gender: A Qualitative Analysis of Trans-Parasocial Relationships with Gen Z Social Media Users", The University of North Dakota Publication	<1%
9	ojs.amhinternational.com Internet Source	<1%