

An Approach to Detecting Fake and Misleading Content on YouTube for the Older
Generation in Bangladesh, Using HCI and Multilingual NLP Models

by

Bivan Shyam Joy

24341084

Meraj Hossen Akib

19101446

Muhammad Abdullah Tanna

19101273

Israt Zahin

19101379

S.M. Zonaed Alam Niloy

24341085

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

Department of Computer Science and Engineering
Brac University
October 2024

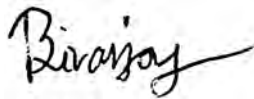
© 2024. Brac University
All rights reserved.

Declaration

It is hereby declared that

1. The thesis submitted is our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:



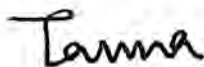
Bivan Shyam Joy

24341084



Meraj Hossen Akib

19101446



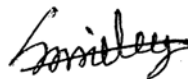
Muhammad Abdullah Tanna

19101273



Israt Zahin

19101379



S.M. Zonaed Alam Niloy

24341085

Approval

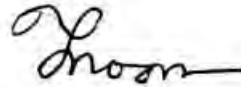
The thesis titled “An Approach to Detecting Fake and Misleading Content on YouTube for the Older Generation in Bangladesh, Using HCI and Multilingual NLP Models” submitted by

1. Bivan Shyam Joy (24341084)
2. Meraj Hossen Akib (19101446)
3. Muhammad Abdullah Tanna (19101273)
4. Israt Zahin (19101379)
5. S.M. Zonaed Alam Niloy (24341085)

of Summer,2024 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on October,2024.

Examining Committee:

Supervisor:



Jannatun Noor

Assistant Professor
Department of Computer Science and Engineering
Brac University

Program Coordinator:
(Member)

Dr. Golam Rabiul Alam

Professor
Department of Computer Science and Engineering
Brac University

Head of Department:

Dr. Sadia Hamid Kazi

Associate Professor
Department of Computer Science and Engineering
Brac University

Abstract

The rapid advancement of digital technologies has dramatically impacted how people access information and connect globally. However, this revolution presents specific challenges, especially for the older generation, who are more prone to fake and misleading content due to their lack of proper digital literacy and awareness. Easier access to the internet and a user-friendly interface for YouTube make it a popular choice amongst the older generation of social media users. However, YouTube's user-friendliness and weak content validation process make it easier for malicious users to deceive viewers and capitalize on misinformation. In this study, we explore the impact of such manipulation on older social media users in Bangladesh, focusing on YouTube. By leveraging Human Computer Interactions (HCI) methodologies, we analyze user behavior to discover how YouTube's deceptive content influences older users' behaviours. Also, we identify vulnerabilities and propose strategies to increase digital safety for this age group.

In this research, we conduct surveys and interviews to identify the types of fake and misleading content that older users encounter and evaluate their ability to identify and protect themselves against it. To address these challenges, we have experimented with NLP models to create a browser extension-based system to enhance digital safety for this age group. Our research aims to improve the YouTube Interface to be more user-friendly while implementing robust and effective content validation processes, ensuring everyone, especially older generations, can navigate YouTube confidently and securely.

Keywords: Human-Computer Interactions (HCI), Misinformations, YouTube, User behavior, Fake video, Older-generation, Digital literacy, Misleading content, Natural Language Processing(NLP) model, Extension.

Acknowledgement

We are grateful and would like to thank all those who have been of help and guidance through the journey of completing this thesis. First, we would like to thank Almighty for his constant support and advice that have enabled us to complete our thesis despite everything. Even if it was only a tiny step, it was a massive step in the direction of our study focus. We needed encouragement and support after every time we failed the experiments. Luckily, there were several people present around us who extended their support towards us. We thank our supervisor, Dr. Jannatun Noor Mukta at BRAC University, for guiding and supporting us throughout this work. Her continuous monitoring, candid critique, sage guidance, enthusiastic support, and active involvement have all combined to make this project successful. Special thanks go out to the participants in this study, whose readiness to share their experiences and insights made this research possible. Their contribution means a lot, and we are thankful for that. This thesis work would not have been possible without the help of our loved ones. We thank our parents and siblings for their understanding, support, and sacrifices in these difficult times.

Table of Contents

Declaration	i
Approval	ii
Abstract	iv
Acknowledgement	v
Table of Contents	vi
List of Figures	ix
List of Tables	xi
Nomenclature	xi
1 Introduction	1
1.1 YouTube and Fake Content	1
1.2 Importance of Extension on YouTube	2
1.3 Problem Formulation and Expected Outcomes	3
1.4 Design Proposal	3
1.5 Our Research Contributions	4
2 Related Works	5
2.1 Older Adults and Social Media Usage	5
2.2 Assessing Credibility	6
2.3 Falling for Phishing Links	7
2.4 Falling for Fake News and Contents	8
2.4.1 Proposed Combative Systems Against Fake Content	11
3 Research Methodology	14
3.1 Data Collection	14
3.2 Questionnaire Design	16
3.3 Data Collection Through Survey Questionnaire	16
3.3.1 Participant Demographics	16
3.3.2 YouTube Usage	17
3.3.3 YouTube Examples	17
3.4 Data Collection Through Semi-structured Interview	18
3.5 Participant Recruitment	20
3.5.1 Snowball Sampling	20

3.5.2	Convenience Sampling	20
3.5.3	Purposive Sampling	21
3.6	Sample Size	21
4	Interview and Survey Findings	23
4.1	Findings from Statistical Analysis	23
4.2	Findings from Thematic Analysis	25
4.2.1	Misconceptions Regarding YouTube Thumbnail vs Content	25
4.2.2	Lack of Knowledge of News/Articles/Published Dates	26
4.2.3	Trust Issues Rooted from Bad Experiences	26
4.2.4	Shyness and Discomfort	27
4.2.5	Limited Knowledge of Social Media Manipulation Techniques	27
4.2.6	Combating The Spread of Misinformation	28
5	System Implementation	29
5.1	Design Proposal A Real-time Extension of “Bangla Shield.”	29
5.2	Architecture Of Our Developed Extension	30
5.2.1	Extension Frontend	31
5.2.2	Pop-up Window	32
5.2.3	Detail Tab	34
5.3	Extension Backend	35
5.4	Dataset Collection	37
5.4.1	Data Labels	37
5.4.2	Initial Dataset	38
5.4.3	Pre-train Model	38
5.4.4	Semi-observed Dataset Collection	38
5.4.5	Final Dataset	39
5.5	Model Training and Testing	40
5.5.1	Natural Language Processing(NLP)	40
5.5.2	Machine Learning	40
5.5.3	Graphs	41
5.5.4	Bangla-BERT Model	42
5.5.5	mBERT Model	46
5.5.6	XLM-R Model	49
5.5.7	Logistic Regression Model	51
5.5.8	SVM Model	54
5.6	Comparison and Findings from Models	56
5.7	Subjective Evaluation	57
6	Extension feedback Survey	59
6.1	Feedback Survey Questionnaire	59
6.2	Findings from Feedback Survey	60
6.3	Thematic Analysis on Feedback	62
6.3.1	YouTube Watch Time vs Fake Video	62
6.3.2	YouTube Watch Frequency vs ”Bangla Shield” Effectiveness	63
6.3.3	Average Confidence vs Age Group	63
6.3.4	Extension Clarity vs. Helpfulness of Features	64

7	Discussion	65
7.1	Challenges Faced by Older Social Media Users	65
7.1.1	Emotional Manipulation Through Clickbait Titles and Thumbnails	65
7.1.2	Limited Digital Literacy and Verification Skills	66
7.1.3	Trust Issues Rooted in Previous Negative Experiences	66
7.1.4	Lack of Awareness About Content Verification Practices	66
7.2	Implementing Technology for Fake Content Detection	66
7.2.1	Development of A Browser Extension for Misinformation Detection	67
7.2.2	Performance of NLP and Machine Learning Models (mBERT, Bangla-BERT, Logistic Regression, SVM)	67
7.2.3	User Feedback on Technological Solutions	67
7.3	Effectiveness of a Fake Video Detection Extension	67
7.3.1	Existing Solutions and Bangla-Shield’s Effectiveness Comparison	68
7.3.2	Limitations of Current Detection Models	69
7.3.3	Potential for Improvement and Future Enhancements	70
8	Limitations and Future work	71
9	Conclusion	73
	Bibliography	73
10	Appendix	80
10.1	Survey Questions	80
10.1.1	A. General Information	80
10.1.2	B. Demographics	80
10.1.3	C. Social Media Usage	81
10.1.4	D. YouTube-Specific Questions	82
10.1.5	F. Phishing and Security Awareness	83
10.1.6	G. Perception of Fake Content	83
10.1.7	H. Example-Based Questions	83
10.1.8	I. Participant Reactions	83
10.2	Interview	85
10.3	Extension Feedback Survey	87

List of Figures

2.1	Illustration of Author’s Example	7
2.2	Change in Republican Margin due to Tweet Bots	10
3.1	Methodology Diagram	15
3.2	YouTube thumbnail regarding dog meat served in restaurant in Bangladesh	17
3.3	YouTube thumbnail example 2 & 3	18
4.1	YouTube usage percentage	24
4.2	Age-Based Reactions to Fake YouTube Thumbnails	24
4.3	Educational Impact on Reactions to Scam YouTube Thumbnails	25
4.4	Fake Thumbnail example	26
4.5	Authentic news.pdf	27
5.2	Extension Icon	30
5.1	Extension Workflow	30
5.3	Extension Flowchart	31
5.4	Color Segment Table	32
5.5	Warning Range: Percentage	33
5.6	Pseudo-code of Pop-up Window	33
5.7	Samples of Pop-up Window	34
5.8	Sample Detail Tab	35
5.9	Server Flowchart	36
5.10	Data Colletion Flowchart	37
5.11	Semi-Observed Data Collector Pseudo-code	39
5.12	Bangla-BERT Training Results	43
5.13	Bangla-BERT Accuracy Vs Epoch	44
5.14	Bangla-BERT Confusion Matrix	44
5.15	Bangla-BERT ROC Curve	45
5.16	Bangla-BERT Learning Curve	45
5.17	mBERT Training Results	46
5.18	mBERT Accuracy vs Epoch	47
5.19	mBERT Confusion Matrix	47
5.20	mBERT ROC Curve	48
5.21	mBERT Learning Curve	48
5.22	XLM-R Training Results	49
5.23	XLM-R Accuracy vs Epoch	50
5.24	XLM-R Confusion Matrix	50
5.25	XLM-R ROC Curve	51
5.26	Logistic Regression Train Result	52

5.27	Logistic Regression Accuracy vs Epoch	52
5.28	Logistic Regression Confusion Matrix	53
5.29	Logistic Regression Learning Curve	53
5.30	SVM Training Result	54
5.31	SVM Accuracy vs Epoch	55
5.32	SVM Confusion Matrix	55
5.33	SVM Learning Curve	56
5.34	Accuracy Metrics Comparison	57
6.1	Willingness to use Browser Extensions	61
6.2	Pop-Up Helpfulness	61
6.3	Willingness to detect fake YouTube videos	62
6.4	YouTube Watch Time vs Fake Video Encounter	63
6.5	YouTube watch Frequency Vs BanglaShield's Effectiveness Among Users	63
6.6	Average Confidence vs. Age Group	64
6.7	Extension clarity vs helpfulness of features	64

List of Tables

3.1	Survey and interview participation by category	21
5.1	Summary of Comments	39
5.2	Performance Metrics of Different Models	57

Chapter 1

Introduction

The rise of digital technologies, especially platforms like YouTube, has surprisingly changed how people access information and connect. While this came up with so many new opportunities for many, it has also created some challenges, particularly for older people. In Bangladesh, an increasing number in the world. However, this group currently has significant risks, such as falling for phishing links, consuming fake or misleading content, and sharing fake content with others, which can lead to financial loss and loss of personal information. Scammers who are spreading this misinformation or fake content- specifically designed their content to trick users into clicking on harmful links or believing false information. Older adults are particularly vulnerable because they don't have that much digital literacy compared to younger generations. Many of them may not have that idea of identifying fake content or scams, making them easy targets. Additionally, factors like intellectual and emotional biases make it harder for them to choose digital content sharply. So, it is easier for them to trust deceptive information, which can sometimes lead to negative consequences. Our goal of this study is to understand better how old people in Bangladesh use YouTube, What kind of content they mostly view, what challenges they face while using this platform, and why they are vulnerable to phishing links and fake content. By using Human-Computer Interaction (HCI) methods, we analyze their behavior and find out the specific areas where they need help. This research will allow us to propose strategies to make a safer place for the older generations on YouTube. We are working on proposing better user interface designs or technology that can detect and flag fake content in real-time. We believe that by focusing on this issue, we contribute a little to making the digital world a safer place for everyone, especially elderly people who deserve the same level of protection and understanding as other groups of people. We hope that this research will not only improve the digital safety of older adults in Bangladesh but also serve as a model for other countries facing similar challenges.

1.1 YouTube and Fake Content

In the modern day and age, when we discuss a video streaming platform for various purposes such as global and national news, entertainment, etc., YouTube is the biggest giant that pops into our minds. Since its creation on February 14, 2005 according to Hosch (2024)[63], YouTube now has more than 2.70 billion active users per month as of October 2024. Per day, 5 billion videos and 1 billion hours of videos

are watched on YouTube by people across the globe. Out of the 63.7% of global social media users, 52% alone are YouTube users (Team, 2024)[64]. This worldwide popularity has unfortunately attracted some people with malicious intent to spread misinformation and content related to various scams. The only other platform with more users than YouTube- Facebook according to Team(2024)[64], has had its own problem with fake content. In 2015, approximately 2% of Facebook’s monthly active users were reported to be fake by Facebook [68], who could easily have spread various misinformation without being detected. While we could not find such statistics for YouTube, we can only assume the problem to be at a similar scale.

Fake content, according to Baptista and Gradim (2022)[52], has five features- it is misinformation online, it contains false statements disassociated from reality, created with an intention to manipulate the masses, is too specific or a product of imagination, attracts the user’s attention through attention-grabbing title or images which provides the maker of the news with higher advertising revenues or success in spreading an ideological propaganda. Fake news is also defined as fabricated misinformation that is spread under the guise of legitimate news (*Research Guides: Fake News and Information Literacy: What Is Fake News?*, 2024) [67]. While known purveyors may share this fake news with ill intent, there may be unwilling purveyors who may unknowingly trust the conveyed false information, and spread it to others on the internet or in society. However, Cambridge dictionary introduces another intention of spreading false news while defining fake news (*Fake News*, 2024) [61]. It may not be only political views or ideology or ill intent that influences creating and spreading false information, rather it can also be performed as a joke. Based on this information, we can define in terms of YouTube that fake content on YouTube is a type of video that contains misinformation that may have been created jokingly or to spread an ideology, for financial gain, or malicious intents- which attracts viewers with attractive or shocking video title or thumbnail image.

1.2 Importance of Extension on YouTube

Browser extensions are small pieces of software that enhance the user experience by providing additional services in the web browser (Agrawal & Srinivas, 2020) [37]. The role of browser extensions becomes crucial as interaction with digital platforms increases day by day. These tools can play a significant role in enhancing the user experience and providing a higher level of security that is not currently available on YouTube.

Currently, YouTube does not provide any built-in tools to detect any fake content or warn users about the videos. By providing an extension that can scan the links, and if it’s suspicious, it will give a warning to adults/users and can help them fall victim to scams. Although YouTube has some measures to take action against misinformation, it’s not enough since, according to their policies, it relies on user reports. Depending on the type of misinformation, such as election-related misinformation or medical misinformation, based on user reports, Google terminates the video and often even the whole channel(*Misinformation Policies - YouTube Help*, 2024) [70]. While it prevents one source of misinformation, it is not a 100% effective method. Also, the YouTube recommendation algorithm sometimes shows the same types of misleading content. By using an extension, users can customize it on their own and can avoid misleading content. So, extension may help to reduce the risk of scams.

1.3 Problem Formulation and Expected Outcomes

Problem Formulation A large number of users are using YouTube to consume information and stay connected with the world daily. It's necessary to make this platform safe and secure. However, this rapid change has come up with significant risks, particularly for old age people who may not be digitally literate enough for which they face problems like financial loss, losing personal information, sharing harmful information, etc. In our research, our main target people are the old generation who face these problems more than the young generation. They are new to this digital world and unfamiliar with digital scams. We aim to find out the reason why old people easily manipulate those scams, and misleading information and develop a good strategy to enhance digital safety and security. Problem Findings are:

1. They have a low level of digital literacy and less knowledge of using digital platforms properly.
2. We identify the weak points for which old age people easily get manipulated by misleading content.
3. By Observing their(old aged people) behavior, we identify the pattern of consumption content, usage, and responses with misleading content.
4. Based on findings and understanding of the targeting problem area, we propose a solution that will improve digital safety and take opinions from them.

Research Questions:

- RQ1: What are the challenges faced by older social media users in detecting fake and misleading content on YouTube?
- RQ2: How can implementing technology help older users detect fake content on YouTube?
- RQ3: Can a fake video detection extension eliminate the barriers older age social media users face?

1.4 Design Proposal

There have been various approaches taken by multiple researchers across the globe to implement a system to fight the spread of misinformation. A design proposed by Reshi and Ali (2023)[69] is a novel fabricated news detection system that uses “contextualized word embeddings generated through ELECTRA-based transformer model as an input to LSTM based deep neural network.” They claim 99.9 percent accuracy in detecting fake news with their system. According to Pérez-Rosas et al. (2018) [20], using 2 novel datasets, they propose a fake news detection system that achieves 76% accuracy.

The difference between our findings on the system designs to combat fake news and our study is that- while there have been systems proposed to combat fake news articles and their spreading, our proposed design is a browser extension that can validate whether a video contains fake and scam content based on the comments provided by previous viewers. The novelty of our research is the lack of such a system to detect misinformation on YouTube.

Expected Outcomes Based on Our Research

1. Understanding the level of their digital literacy, especially on YouTube, and approaching them with a user-friendly design so that they can easily explore YouTube.
2. We will highlight the strategies and tricks that scammers usually use to scam them and help them understand what kind of content they need to avoid.
3. Based on our observation, we will come out with a solution where they can easily identify the misleading content and feel safe while using YouTube.
4. Our goal is not only to approach a solution but we want to build a real-time solution so that it can improve digital safety.

1.5 Our Research Contributions

This research makes several important contributions to understanding and improving digital safety among older adults in Bangladesh, especially on YouTube. We hereby make several important contributions to the fields of HCI and digital literacy. First, we provide a complete data analysis of how fake content specifically targets older adults in Bangladesh and what kind of strategy they use to manipulate old generation people easily. By focusing on this group, we observed that not only this generation has fallen into these kinds of traps but they are at high risk of manipulation in the digital platforms. Secondly, we propose a set of design interventions that are made specially to find out those challenges faced by older users in Bangladesh. These designs are informed by HCI principles and user-friendly approaches that prioritize simplicity. Once this has been addressed, we underline some targeted interventions that may identify those factors that make them vulnerable in this aspect. Third, the research contributes to a deeper understanding of the behavioral patterns of older social media users, identifying key vulnerabilities that make them more susceptible to manipulation. By examining these behaviors through the lens of HCI, the research offers new insights into how older users interact with digital platforms and what can be done to improve their safety. Finally, this thesis offers practical strategies for enhancing digital safety, ranging from educational tools to personalized content filters. These strategies are designed not only to protect older users from manipulation but also to empower them with the knowledge and tools they need to navigate social media platforms more securely. The contributions made by this research have the potential to inform both policy and practice, offering valuable guidance for developers, policymakers, and educators working to improve digital safety for older adults in Bangladesh and beyond.

Chapter 2

Related Works

This part aims to review previous relevant work on how old people from particularly different places fall for fake content, misinformation, and phishing links through social media platforms such as Facebook and YouTube, with the idea and context of Human-Computer Interaction (HCI). Moreover, we also show the various processes and different techniques used to analyze and reach our main output. On the other hand, some of the challenges we face in conducting the research include convincing an older age group of their beliefs, different genders, psychological behavior during face-to-face interviews with the interviewer, and the collection of qualitative and quantitative data. The variety of devices used to access, lack of knowledge of how to use social media platforms or any tech device, and inability to judge the credibility of any content complicates the process.

2.1 Older Adults and Social Media Usage

In their study, Bell et al. (2013) [9] conducted a survey in a small range demonstrating social media users focusing on only Facebook from the ages of 52 to 92 in Atlanta, where the response was conducted among 142 participants and the response rate was 54 percent. This survey was conducted using some methods such as statistics, such as the participant's gender and age, loneliness scale (social), satisfaction scale, confidence with technology, SNS use, etc. Moreover, in this approach, gender and age, race, or income were not significant predictors. Income also played a role in locating users, as users earning more than 30,000 USD per year were more likely to use Facebook than non-users earning less than 30,000 USD per year. Whereas, on social media, Facebook users enjoyed more social satisfaction. Nevertheless, the most significant attitude among the users and non-users was their attitude towards technology. To conclude, applying different parameters, the authors found that social media leads to higher social satisfaction, tighter family bonds, and a strong chain in the community, even though most older adults don't have a clearer idea of Facebook's privacy and security issues, communication preferences, and rigid and authentic contents that they are viewing. Two other prior studies that were conducted in India by Shahid et al. (2022) [56] and Shahid et al. (2022) [55] were performed based on mainly two demographics- rural residents and urban residents. While the authors performed the studies on people up to 65 years old, the age group was not the focal point of their research. Their studies were more dependent on the region from which their participants came.

2.2 Assessing Credibility

A study by Seo et al. (2020) [50] was conducted among lower-income African-American older adults who are vulnerable to misinformation and get manipulated without checking the credibility of social media. The authors conducted face-to-face interviews and analyzed the quantitative data of the survey. As a result, the authors found out how these old adults' digital media use, demographics, self-efficacy, and participation in topics were related to credibility assessment. Education, socioeconomic status, and topic involvement are the two main reasons why old adults fall for misinformation and get manipulated. Moreover, this research work shows older adults of African Americans are rarely skilled in online information and tech usage resulting in not having a proper knowledge in judging the quality of content and information retrieved. 125 low-income African Americans completed the survey, and 15 people were interviewed face to face, where women were in huge numbers falling for misinformation. In another paper by Quan-Haase and Elueze (2018) [26], researchers investigated the limitations of privacy issues and obstacles regarding this demographic's use of social networking. Moreover, the types of social media privacy concerns that older adults have, as well as the methods they use to decrease these concerns, by conducting in-depth interviews with 40 older people in East York, Toronto, Canada, who use and do not use social media. Although elderly people (65+) utilize a variety of online media, they have taken their time embracing social media in particular. Moreover, researchers discovered that older adults who used social media and those who did not had comparable privacy concerns. The most often expressed concern was for data theft and abuse of personal information. Older adults who don't utilize social media secure themselves by limiting the information they share. In another study, Shahid et al. (2022)[55], the participants were shown different types of fake videos, such as misleading and out-of-context videos, digitally modified videos, deepfake videos, etc., and were asked to identify these videos as fake or real. Nearly 1/3rd of the participants believed all the presented videos to be real, nearly 1/4th believed the real video was fake as well, and only one participant correctly identified all the videos. Among these participants, some had no idea about the existence of fake videos and the possibility of tampering with videos. 3/4th of the participants could not classify the meaning of deepfake but were aware of the existence of deepfake videos. In terms of identifying the fake videos, the participants mentioned some key factors such as the usage of foreign language, videos focused on defamation, misleading titles such as "If you do X, Y will happen." with a religious undertone, low resolution, discovering audio manipulation, etc. In another study, Shahid et al. (2022)[56], the authors focus on the participants' ability to identify the differences between credible and fake posts on Facebook and the possibility of them sharing the posts based on their beliefs. Here, they grouped their findings based on urban residents vs rural residents. Using six posts- 2 of which were fake, the authors established a median trust rating of 4 or higher as the margin of the participants passing. Compared to urban residents who averaged a rating of 4, the rural residents averaged 3 out of 5. While urban residents were willing to share 64 percent of the posts and most of it (32 percent) to a public audience, the rural residents were willing to share a lesser amount of posts (35 percent) and most of it would be with their friends and family (25 percent). In a research article by Ferrara (2015) [16], the author discusses how people are susceptible

to different sorts of cyber abuse and manipulation tactics and how this manipulative misinformation can cause real harm. At first, the author explains with examples how manipulative misinformation can spread panic and fear and cause real-world harm. For example, figure 2.1 is a graph that shows a 136 billion USD loss from a misinformation tweet. The author brings several different examples that have been the reason for widespread panic and monetary loss. Afterward, he explains the illustrations of stated examples and gives his suggestive input on how to combat this issue.



Figure 2.1: Illustration of Author's Example

Research by Figueiredo et al. (2014) [12], focuses on what exactly drives the popularity of information on social media (specifically YouTube). By asking several users about their choice of different types of videos on the site, they understand the extent to which content by itself determines a video's popularity. They found that in most evaluations, users could not reach a consensus on which video had better content as their perceptions tend to be very subjective. However, when consensus was reached, the video with preferred content almost always achieved greater popularity on YouTube, highlighting the importance of content in driving information popularity on social media. This kind of information helps us detect the downsides of social media sites and how they can focus their users or drive their audience towards content, whether it's trustworthy or not.

2.3 Falling for Phishing Links

Oladoyinbo (2024) [71] studies the effects of phishing schemes that target American people between the ages of 50 and 80. The study looks into how susceptible senior citizens are to internet phishing schemes that take advantage of their unfamiliarity with digital security protocols. Since older people are frequently more vulnerable to dishonest tactics, the study highlights the financial, psychological, and emotional effects of these scams on this demographic. In order to safeguard this group, it also addresses the wider social and economic ramifications of such fraudulent activity and calls for stronger security measures and educational initiatives. The results imply that lowering the risks connected with phishing scams can be achieved by

raising awareness and offering senior citizens specialized tools. In this paper Seng et al. (2019) [33], authors use a simulated interface and a think-aloud protocol to address the challenge of falling for phishing links via posts and personal messages, which ultimately result in the user having a malware attack on their device, seeing manipulated advertisements, having their personal information leaked, or even becoming a victim of cyber-crime. Furthermore, the authors hope to make further progress in understanding the impact of a user's decision to click by conducting a vignette study that will allow users to think deeply about real-life scenarios. To illustrate, authors conducted this study in a LAN setting with the consent of the participants in the survey, resulting in a total of 48 vignettes considering all combinations of values where the likelihood of clicking on those links was based on relationship, place, post type, marketing type, and ads generated. Inbox and tagged posts were also studied, as was the user's clicking reaction to the content. In this vignette study, researchers looked into the relationship between the attributes of a social media post, such as who created it, where it is placed and its type, and the probability that users would click on it. Understanding this information is crucial for developing user safety systems because it is likely used by attackers attempting to phish Facebook users. The paper by Wu et al. (2011) [8] is about the detection of phishing by using genetic algorithms. By using this algorithm, we can solve the problem of phishing. It's a rule-based system where we can differentiate phishing links and legitimate links. Any legitimate website owner can use this to protect his webpage from Phishing links. This algorithm sets some ruleset which only matches with phishing links. So, if any phishing links are detected, it will report the link immediately. So, it can be a good solution for anti-phishing. In a research by Li et al. (2020) [62], the authors have attempted to identify susceptibility to phishing links based on different demographics such as age, gender, employment, etc. The test subjects were 6938 faculty and staff at the researchers' university to identify respectable users' traits and characteristics to understand who and why they are more prone to getting phished than others. Over three weeks, the authors sent several different phishing emails on different topics to the people to observe what sort of topic attracts what type of people and collect and analyze data based on the results.

2.4 Falling for Fake News and Contents

A study by Hayes (2023) [68] elaborates that cyber criminals run rampant across every social network today. Poor social media security practices put their brands, customers, executives, and entire organizations at risk. Facebook reported that for 2015, up to 2% of its monthly average users were fake. Twitter estimates 5%, and LinkedIn openly admitted, "We don't have a reliable system for identifying and counting duplicate or fraudulent accounts". Social media sites are now a treasure trove for cybercriminals. LinkedIn was a key tool for reconnaissance of the Anthem Health 2015 breach and its 80 million stolen records. Twitter was an integral component of an innovative malware exploit dubbed "Hammertoss". All of this shows us why social media abusers can redirect inexperienced users towards a dark path and abuse their behavioral outcome. While social media sites may not create completely new cyber threats, they do substantially amplify the risk of existing ones.

A research by Geeng et al. (2020) [40] focuses on users' Facebook or Twitter feeds

and their ability to judge and detect any sort of fake news posted by someone familiar. A partially structured interview was conducted to see if someone using these platforms for strictly news or reports can detect any fake news, which was made to appear using a third-party extension unbeknownst to the test users. It seems that people don't tend to properly commit to fact-checking and verifying posts/news from less credible sources. Also, many people take the contents of trusted posters at face value, even though the poster's creator might have ill intentions in a specific instance to deceive the people.

In their paper by Almaliki (2019) [28], the author says that gamification methods can be used for identifying misinformation. But while implementing, the model needs to consider a variety of sentiments and preferences as peoples' mentality varies. According to their data, 67 percent of users contributed to misinformation, while 94 percent of users believe they witnessed false information. In their opinion, misinformation can cause serious negative emotions, misunderstanding, and anxiety amongst the users, and some lead to cybercrimes. In the end, people's reaction to this gamification method varies because of different factors. Based in Bosnia and Herzegovina, Trninić et al. (2021) [51] studied the people mainly young and middle-aged generation to identify how much they can perceive fake and manipulative content- how much they can verify or recognize such content on digital platforms. For conducting this research, the authors selected "focus groups" with a specific number of digital media user participants (no less than 6 and no more than 12). The target generations were identified in two age groups- young (18-34 years old) and middle-aged (35-65 years old). With 24 participants in different age groups, the researchers held sessions in which they discussed with the participants in detail their perception of fake news. They also collected datasets by making the participants fill up certain questionnaires. The authors used a "qualitative thematic analysis method" to analyze these datasets to conclude this research.

A study by Shao et al. (2017) [22] conducted in 2017 discusses how misinformation and fake news have been spread on one of the major social media websites- Twitter, using bots. The data used for this research were 400 thousand claims spread throughout 14 million messages at the time of the 2016 US Presidential Election. The research focuses on finding out how people were manipulated by software-driven AI bots tweeting falsified information during the election. To detect these bots and their manipulative tweets, the researchers used two self-developed tools that identified the bots and tracked how far the claims were spreading. Using independent fact-checkers and misinformation publishers, they identified the amount of fake news tweeted by these bots. After collecting these data, they analyzed how these manipulation tactics fared in the election. For example, figure 2.2 in their paper shows how the republican margin changed due to tweet bots.

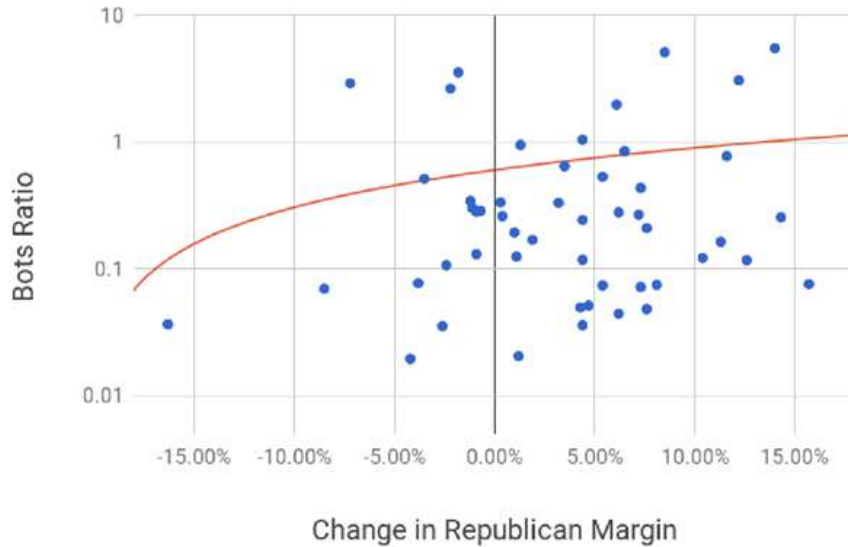


Figure 2.2: Change in Republican Margin due to Tweet Bots

A research by Alkış and Temizel (2015) [14] helps us see that persuasion can play an important role in driving people or their behavior towards something, which can be in a good or bad way. These persuasions are taking place in different contexts, such as in online commerce, fundraisers, ads, or even health-related info systems, and the effect of this varies from individual; their response can be varied. This research helps identify which personality traits are significant in determining individuals' susceptibilities to influence strategies of Cialdini [3]. Individuals who are not that into openness traits are more likely to be driven towards authoritarianism & vice versa. These people are targeted based on their approach, agreeableness, conscientiousness, or something like neuroticism. Then, different strategies are inclined. Individuals with neuroticism and extraversion traits are more inclined to use scarcity strategy than others. This study shows and helps us to learn about persuasion and its applications, which come out best when they are tailored to individuals. These sorts of personality traits & persuasion strategies can be driven towards social media exploits. A study of Moravec et al. (2018) [25] helps us observe firsthand how social media can be weaponized and used in a dark sense. It's done using behavioral EEG data from social media users, how their judgment, cognition, etc. are getting affected. It's clear that these users are able to detect some obvious absurd news but only 17% of their participants could detect fake news. Most users (84%) believe they can detect fake news on social media Barthel (2016) [17]. First, This paper focuses on whether users can detect fake news on social media. The second focus is on two factors that may influence users' beliefs: confirmation bias and fact-checking. This study uses neurophysiological responses measured by EEG as an indicator of cognitive activity. The results are based on behavioral and neurophysiological. Proper Implications for practice & research are there with limitations.

From the mentioned papers and discussion, it can be observed that to complete the work process, most of the authors have come up with ideas like analyzing demographics and online surveys via Google Forms. On the other hand, to analyze the credibility of phishing, link face-to-face interviews showing fake and original content at a time and let the participants judge. Moreover, to analyze the psychology in

terms of spreading fake news and misinformation, some researchers went for clicking on response and reaction time to misinformation about how old people are getting manipulated. Some used eye-tracking technology to show how fake content on social media platforms caught the attention of more and more old people. Furthermore, the process and the surveys conducted have fewer participant responses and some also do not come as a rigid output as face-to-face interviews in the Laboratory might influence the normal behavior of the participants and they may not act as they would in their home environment.

2.4.1 Proposed Combative Systems Against Fake Content

A study by Ruchansky et al. (2017) [21] focuses on fake news which has the power to change people's perceptions. Automatically identifying fake news is a significant, difficult, and poorly understood problem due to the high stakes involved. Nevertheless, three characteristics of fake news are generally acknowledged: the content of an article, how users respond to it, and the users who promote it from the source. So, they present a model that incorporates all three characteristics. They propose a model that is called CSI, which is motivated by the three characteristics and is made up of three modules: Score, integrate, and capture. Based on the response and text, the first module captures the temporal pattern of user activity on a particular article by employing a Recurrent Neural Network. The second module integrates with the third module to determine whether an article is fake or not by learning the source characteristics from user behavior. Real-world data analysis shows that CSI is more accurate than existing models and extracts meaningful latent representations of users and articles.

According to the work of Ngada and Haskins (2020) [41], a major challenge in implementing any feature to combat fake content detection is the complexity of human dialect. The authors used a pre-existing dataset [60] with a fake-to-real news ratio of 23481:21417 and six machine learning algorithms to separate fake news from reliable news- the two types of separation they created among contents to verify (Fake-and-real-news-dataset, 2024). Finally, they used the Confusion Matrix model for performance analysis and 10-fold cross-validation for proper validation of their results. The dataset was split in an 80-20 ratio for training and testing. In terms of accuracy, SVM achieved the highest 99.4% accuracy.

In their work, Abdulrahman and Baykara (2020) [36] have divided their work into 3 stages- pre-processing, extract features, and classifiers. The authors used a dataset containing textual data, which they cleaned and preprocessed by removing non-English, removing HTML tags, and applying the stopword technique. For extracting features, in which they converted the text data into vectors 0 and 1, they used multiple vectorizers such as TF-IDF, N-gram level, character level, and count vectorizer. Afterward, they tokenized these vectors and finally, in the last stage, ran 2 types of classifiers- Machine Learning classifiers and Deep Learning classifiers. With the TF-IDF extraction technique, they achieved a 100% rating accuracy using the AdaBoost classifier.

In the work of Ahmad et al. (2020) [38], the datasets used are publicly available at Kaggle by Lifferth (2018) [66], (Fake News Detection, 2017)[43] and ISOT Fake News

Dataset by Ahmed et al. (2017) [24]. While preprocessing the data, they removed unwanted redundancies such as the name of the author, URL, categories, date, articles with no body text or having less than 20 words, etc. After cleaning the data, using LIWC2015, they extracted various linguistic features such as the percentage of words conveying positive or negative emotions, informal language, grammar, stop words, etc. Afterward, a 70/30 split was applied to divide the data for the training and testing set. The novelty of these authors' research is that they applied various ensemble techniques to evaluate the performance over multiple datasets. Two voting classifiers were used- the first one comprising logistic regression, KNN, and random forest, and the second one comprising logistic regression, SVM, and CART. On the ISOT dataset, they achieved 99% accuracy with the random forest algorithm and Perez-LSVM. From the first dataset of Kaggle, 81.5% accuracy with ensemble learners could be achieved, but in the second one, Perez-LSVM was the winner with 96% accuracy. With a 91% accuracy rate, the random forest algorithm worked better on the overall dataset.

Shaikh and Patil (2020) [44] have used text data retrieved from news headlines and articles- which they then cleaned using stop words removal through Natural Language Toolkit (NLTK) library, stemming and punctuation removal. Using Term Frequency (TF) and Term Frequency-Inverted Document Frequency (TF-IDF), they performed feature extraction. Lastly, they implemented three classifier algorithms- Support Vector Machines (SVM), Naïve Bayes, and Passive Aggressive Classifier. Their training and testing split of the dataset was 80%-20%. From their results, they propose an SVM approach, which yields them 95.05% accuracy rate.

In this study by Waikhom and Goswami (2019) [34], a publicly available LIAR dataset has been used to train a model to detect fake news. Containing 13 different features across 12.8k samples, this dataset was pre-processed by dropping all NULL value rows, converting text-based data to N-gram, and then to TF-IDF vectors. Bag of words was also used to extract features. After applying label encoding on categorical columns, some columns were One hot encoded to ensure better accuracy. The numerical columns were scaled using the Min-Max algorithm. The dataset was split in an 80:20 ratio for training and testing purposes. After binary classification, Bagging and AdaBoost classifiers achieved 70% accuracy in precision.

In their work, Kaliyar et al. (2019) [30] have used the Kaggle fake news dataset. Using TF-IDF, Cosine Similarity, Hand Selected, Word Overlap, Polarity, and Refuting features, they performed feature selection. Among multiple classifiers used in similar research stated before, these authors also used Gradient Boosting, since this classifier trains a few models gradually and successively. Their research yields an 86% accuracy rate using Gradient Boosting. This accuracy is defined as the performance of the researchers' learning algorithm for correct predictions made from 11651 instances available in their dataset.

In their research, Wynne and Wint (2019) [35] have used N-gram features in order to detect fake news. During the data pre-processing stage, they cleaned the raw text data from news headlines and articles by eliminating the punctuation, capital to small letter conversion, removal of stop words, and stemming. Then they tokenized

the data to use N-gram. Both words and characters were used as n-gram features. TF and TF-IDF techniques were used during feature extraction. The researchers have used two Kaggle datasets to train and test- a 'real or fake news' dataset with 6256 articles with headlines and a 'fake news detection' dataset containing 4009 articles. Gradient boosting proved to be the best classifier to generate outcomes for their research.

The work of Sharma et al. (2021) [45] proposes a system with functions divided into 3 parts. The first part, which is static- functions on a machine-learning classifier. Here, the authors have used 4 classifiers to train the model and chose a Logistic Regression classifier to identify the fake news. The second part, which is dynamic-takes a keyword or text from the user and searches online to calculate the truth probability of the news. The third part provides the authenticity to the user. The authors used the LIAR dataset and REAL_OR_FAKE.CSV to train the model. Data was cleaned during pre-processing by deleting stop words, removing punctuations, and performing tokenization. The authors used 3-fold cross-validation system where 67% of the data was used for training and 33% of the data was used for testing. While logistic regression provides 65% initial accuracy, the authors have improved that to 75% by applying grid search parameter optimization. In this research [46] the authors tried to find the effects of propaganda on social media. They used Twitter API to extract posts with keywords and sort them. They then used an ML support-vector machine and distributed the data into two sections 1. propaganda 2. Non-propaganda. They found that SVM has 69.84% Accuracy with F1-Score 0.81 for the Non-Propagandist class and 0.58 for the Propaganda Class. Logistic Regression demonstrated 68.76% Accuracy with F1-Score 0.78 for the Non-Propagandist class and 0.58 for the Propaganda Class. They also found that SVM beats all other traditional machine learning algorithms by having a recall of 0.99 and 0.54 with Precision of 0.69 and 0.71 separately of two classes. Their additional finding was that propaganda posts have longer sentences compared to non-propaganda posts.

Chapter 3

Research Methodology

3.1 Data Collection

According to research by Vaswani et al.(2023) [59], OA are motivated to learn by internal considerations like fulfillment and personal growth. Regardless of their level of digital competency, five of the youngest OA (ages 74–85) said they wanted to keep learning in order to feel proud and accomplished, while five of the oldest OA (ages 86–91) said they had no interest in using ICT because they found it difficult to "learn anything new" or because they could complete daily tasks without ICT by using friends or family as a stand-in. However, all OA acknowledged that digital literacy has quickly changed from being a choice to being essential for surviving in contemporary life.

To understand elderly people's approach, behavioral changes, and valuable opinions, we decided to divide them into age groups, starting at the age of 40. As a result, we conducted a quantitative survey, a semi-structured interview, an in-depth interview, a proxy interview, and focus group discussions to gather rigid data for our research, as illustrated in Figure 3.1. Selecting and identifying the correct population (sample) for the study is a crucial step in the qualitative research process as it allows data to be gathered from that group. The population under study in qualitative research is almost usually human, however, there are a few exceptions that will be covered in this section. In qualitative research, a human being is frequently referred to as a participant (or occasionally a subject). A selected group that is typically representative of a larger population is called a population sample. This chapter focuses on the methods used in qualitative research for both data collecting and sampling.

In addition, sampling methods [10] like snowball sampling, convenience sampling, and purposive sampling were used as qualitative research techniques. For this reason, a large population becomes our target audience, and therefore, greater statistical power and ability to gather information come with a survey. But before the survey, we designed a lenient set of questionnaires keeping the older age groups in mind, where we used both Bangla and English to properly understand their views on YouTube videos, news, and links. To start with, we used different platforms to reach out to people interested in participating in our qualitative survey, where we allowed other people to fill out the online survey on behalf of their known ones

who belonged to our targeted age groups. Moreover, we observed from existing literature that many people from our target audience were using smartphones or devices with internet connectivity but were not efficient enough in using Google Forms.

Afterwards, it was time to recruit participants, so we reached out to X, Reddit, email, Facebook student community, research groups, and online discussion forums and invited users from the internet to participate in our survey. We offered the participants to participate in our survey, and as remuneration, participants were given online currency like torrent seed bonus points, and a few were given mobile top up. Furthermore, for semistructured interviews, we went to Dhanmondi Lake and a housing society in Dhanmondi to conduct interviews with their consent and keep the audio recording to keep it on record where 17 interviews were conducted, and for that, we prepared a questionnaire too with some demographic information and probing questions highlighting our main focus YouTube. The examples shown were relevant to the interviewees as well. They expressed their initial reaction to the examples and gave valuable opinions and suggestions on preventing phishing links and fake videos circulating over YouTube.

As a part of snowball sampling, where we used or knew one’s friends group, family, and university community, we recruited interviewers on our behalf, taking proxy interviews with their parents, relatives, or known ones who match our target audience, and we provided the interviewers with the same semi-structured interview questionnaire we used during face to face interviews.

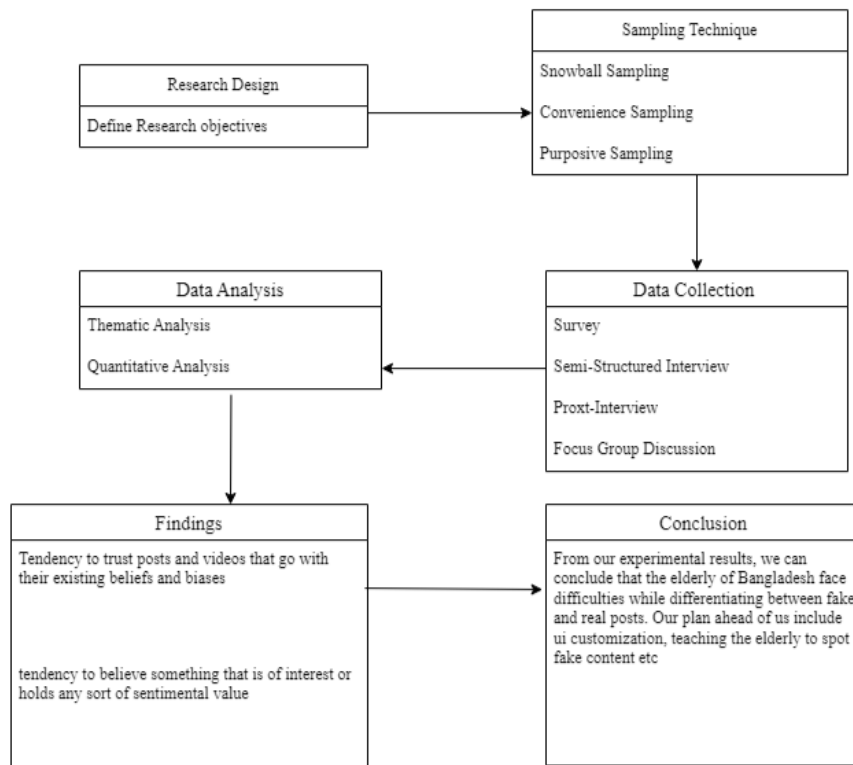


Figure 3.1: Methodology Diagram

3.2 Questionnaire Design

The questionnaire itself is a valuable tool for collecting data for research work. Because a questionnaire provides standard questions and response options, all participants receive the same set of questions. Questionnaires are nicely fitted for gathering quantitative data to analyze it easily and draw statistical conclusions, as they are a cost-effective and efficient way to collect data from a large number of participants simultaneously. Data collected through questionnaires can be easily entered into statistical software for analysis, making it relatively straightforward to generate graphs, tables, and statistical tests. Moreover, we feel that through a questionnaire, sensitive topics and the opinions of participants remain rigid and appropriate as their responses can be delivered anonymously. Overall, questionnaires play a crucial role in survey research by providing a systematic and efficient way to collect data, enabling researchers to make informed decisions, draw conclusions, and contribute to our understanding of various phenomena.

3.3 Data Collection Through Survey Questionnaire

Keeping our research topic and target audience in mind, we had to go through many related works done before and keep the survey questionnaire very simple and understandable. In addition, older adults may vary in digital literacy level and cognitive abilities, so we decided to set up the questionnaire in both English and Bangla, as Bangla is the native language of Bangladesh, and our target audience is from this country. We tried to understand their psychology of having an interest in the topics that they often search for or like to see and put them as examples in the questionnaire, such as sports, religion, politics, and worldwide current affairs. We divided the questionnaire into four parts. i. Participant Demographics ii. YouTube Usage iii. Examples and Conclusion

3.3.1 Participant Demographics

As mentioned earlier, starting at age 40. 40-50, 51-60, 61-70, and above 70—these are the age groups we divided. In this part, we collected the participants' age, gender, highest level of education, and the areas where they are living for a better understanding and to build a relationship based on the facts about whether these factors were affecting the participants or not. In addition, social media usage, the most trusted platform as a news source, was also taken as input on the questionnaire. Demographic information is often used to establish connections during the data analysis period. These are the core factors helpful in our research work. Understanding participant demographics aids in the generalizability of survey findings. By knowing the characteristics of the sample, researchers can better determine to which populations or groups their results are applicable. This helps in making broader claims about the findings' relevance beyond the survey sample. Furthermore, demographic data collection is essential for ensuring ethical research practices, such as avoiding bias and protecting vulnerable populations. Researchers can use this information to ensure they are treating all participants fairly and ethically.

3.3.2 YouTube Usage

This particular section itself is vital because nowadays YouTube is booming in our country, and older adults have found their form of entertainment through this platform. Print media and electronic media, such as newspapers and TV channels, are now available on platforms like YouTube.

For this section of the questionnaire, we tried to check their knowledge of the platform and also if they are aware of the YouTube policies and how the YouTube recommendation system works, or if they are manipulating them into certain topics of videos or not. Firstly, we asked about the hours they spend regularly and the genres of videos that interest them. Moving forward, we asked them about the news, facts, or content shown in the YouTube video and how they evaluated the credibility of those contents. Moreover, the possibility of spreading false news or misleading information where older adults might get manipulated was also our focus. Lastly, the YouTube thumbnail and the inside content of the video also play a role in manipulation. We tried to collect the responses by asking them how they deal with such misleading videos if they skip, ignore, or report the videos as false information spread. Open-ended questions made this section more communicative with the participants as well.

3.3.3 YouTube Examples

Apart from text-based questions, multimedia examples are the best way to gather the data. Here, we introduced four YouTube video thumbnails and asked the participants a few questions.

To give examples regarding YouTube, only shorts and videos were the medium, but we chose the thumbnails of some videos and wanted to examine their initial reaction and eagerness to watch the video or if they had already seen the types of examples we set up in the questionnaire. If they had already seen it, then we asked if they verified the source or who shared those YouTube videos with them.

Firstly, a news video from a popular TV channel regarding puppy theft indicates that the meat is to be used in restaurants. From recent incidents, we found this video- the thumbnail of which is shown in Figure 3.2- to be relevant to the participants.



Figure 3.2: YouTube thumbnail regarding dog meat served in restaurant in Bangladesh

Secondly, YouTube BD, meaning the Bangladeshi version of YouTube, is filled with different videos of earning money from home (keywords: new site 2024, new earning

app, play games and earn money, easy money apps, bKash transfer within minutes). For this reason, we picked two examples shown in 3.3 where people were giving unbelievable reach to the content, which they perceived to be true. Even phishing links were given in the description box of the YouTube videos. We wanted to understand how older adults react to these online earning apps or if they are already familiar with these kinds of videos. Moreover, we set up another example titled "Lifetime Free YouTube without Mobile Data" and asked our participants, by seeing the thumbnail, what would be their first thoughts and if they would share it with others or not. Reasons and whom they would share were also asked in the questionnaire. We, as a group, did not declare or indicate the truthfulness of the content. As a result, we were able to reach our goal from the example section.



Figure 3.3: YouTube thumbnail example 2 & 3

Lastly, the whole survey questionnaire was designed very carefully so that the participants did not lose interest in filling it up. Also, simple language usage, decent text size, standard examples, and to-the-point questions were set to design the questionnaire.

3.4 Data Collection Through Semi-structured Interview

To begin with, semi-structured interviews[6] in HCI research provide a rich, flexible, and user-focused approach to understanding the complex interactions between people and technology. This method allows researchers to gain deep insights into user

experiences, preferences, and contexts, which are critical for designing effective and user-friendly technology solutions. A semi-structured interview is a part of the qualitative research method where the interviewer prepares a set of questions or topics related to the study beforehand which serves as a guide to ensure the targeted or relevant topic is fully covered in the interview. Although having a set of questions prepared is not necessary to maintain the sequence, interviewers can diverge from it by asking follow-up questions and exploring the topic more. Furthermore, it is always participant-centered where participants are encouraged to express their own opinions easily. As per the recording of the data we have used both audio recording and taking notes.

To begin with, we thought of conducting a face-to-face semi-structured interview. As we were distributing and recruiting participants for our survey and interview. We also offered from our research group community if anyone is interested in interviewing participants on behalf of us someone he or she knows having our targeted age group from forty to above seventy. As a result, we recorded the audio of in-depth depth interviews and got the note-downs and handouts from the proxy interview part.

Firstly, we focused on how simply we conduct this interview, keeping the interview formal but also engaging the participants in the conversation. To understand our topic, participants of the interview must know some terms like fake posts and fake videos. Starting with a greetings message we humbly took the permission to record the interview throughout. Starting with some basic questions like their social media usage, and how many hours they spend daily, and then gradually we moved.

Secondly, we had to test the knowledge of the participants if they knew about fake posts, fake videos, and misleading information. If they are not aware we would make them familiar by defining the terms mentioned. Moreover, we set up some questions regarding their experiences if they have encountered fake videos, posts, or links.

Thirdly, we prepared and set up two examples from the YouTube platform to record their instant reaction and hear their opinion or the facts about why they would believe or disbelieve the YouTube examples.

Fourthly, after showing them the examples, we would ask for their opinion and the truthfulness of the news or videos, whether they were rigid, fake, edited, modified, or misguided. We also set up some probing questions if the participant gets nervous or confused. Fifthly, we would ask them, according to their answer (right or wrong), if they would have shared the content or what clues made them think of its truthfulness.

Sixth, as a part of the questionnaire, we set an example of a phishing link where the uploader of the YouTube video creates a clone website and betting apps and tittles it to earn free money, which can make people fall for that phishing link, resulting in a cyber victim.

Lastly, we set the questionnaire to determine if the posts and videos shown could be beneficial to them or harmful to them or their willingness to share the content to create awareness or share the content in favor of the uploader. Moreover, they should be more conscious of preventing this misleading and fake content, giving them the freedom to answer. In the last part, we take some demographic information like

age and educational qualification and end the interview by showing gratitude to the older adults.

3.5 Participant Recruitment

Participant recruitment plays a vital role in research work where sampling methods are used. Qualitative sampling methods include convenience sampling, purposive sampling, snowball sampling, and theoretical sampling. To be precise, the primary purpose of sampling is the selection of suitable populations so that the focus of the study can be appropriately researched. In qualitative research, an effective sample selection process is very important because inappropriate procedures may affect the outcomes of a study.

3.5.1 Snowball Sampling

Snowball sampling, which is also called "chain referral" or "networking" sampling, is when a researcher starts by getting information from one or a few people and then relies on those people to put them in touch with other people, who could be friends, family, coworkers, or other important contacts. This results in recruiting a "chain of participants." [31] This method of sampling works best when the sample is made up of people who are on the outside or who are stigmatized. It can also be used to find and gather people from "hidden populations," or groups that are hard for researchers to reach with other sampling methods.

Firstly, we reached out to our known and close ones, including friends, family, and neighbors, and requested that they distribute the online survey questionnaire on our behalf. Moreover, we tried to recruit participants through our connections and this sampling method can be considered snowball sampling. As our targeted audience is different groups of older adults, it's quite complex to gather people, and to reach a larger number of participants, we used our 'networking' as a form of data collection. Furthermore, we reached out and distributed the survey questionnaire via social media platforms as well.

3.5.2 Convenience Sampling

This is the most common type of qualitative sampling. It happens when people are asked to take part in a study because they are willing, able, and able to receive the information. A quick and easy way to get the sample size needed for the study is to use convenience sampling. This type of sampling allows the researchers to select more representative samples and generalize the results [53] To begin with, we announced Facebook Messenger, posted in the YouTube community and X as well about our research topic and for remuneration, we made the announcement that recruited participants will get twenty taka mobile top-ups each.

Secondly, apart from reaching a larger population, we decided to reach out to specific communities. As a result, we posted on the timeline of an online forum an announcement about recruiting participants to participate in our survey or interview older adults on behalf of our research group.

Thirdly, we went to Dhanmondi Lake in order to make this survey more fruitful, as older adults are often seen walking or jogging around the area. With consent, we

took a few minutes to educate them on our research purpose, and convinced them to participate in our semi-structured interview.

3.5.3 Purposive Sampling

This is another common sampling method in which people are asked to take part based on factors that have already been chosen to be relevant to the research question. Purposive sampling, which is also sometimes called "judgment sampling," is meant to give researchers cases with a lot of information to look into in more depth. This is because participants are people who are qualified to give experts the information they need because they have the right status, experience, or knowledge. Getting family, friends, and other older people we know to join by using personal ties. Going to a housing society and leading a talk with a focus group. We only wanted to reach a certain group of people in that living society. This kind of selection is called "purposeful sampling."

3.6 Sample Size

We surveyed 42 male and 23 female participants, and women representation was slightly dominant in the interviews (4 male and 6 female). Participants of both genders were represented equally in focus group discussions but not in proxy interviews (2 male, one female). Participants aged 40-50 were the most represented participants in the surveys (43) and also in focus groups (2), followed by 50-60 age group participants in surveys (16) and 1 in focus groups. However, only five people from the 60-70 age group participated in the survey. Most of our interviewees were from the capital city (7), followed by residents in the city area (3). Undergraduates (41) were the most surveyed group, followed by postgraduates (17). We found only

Category	Sub category	Survey	Interview	Focus Group	Proxy Interview
Gender	Male	42	4	2	2
	Female	23	6	2	1
Age	40-50	43	3	2	0
	50-60	16	6	1	0
	60-70	5	1	1	2
	70+	0	0	0	1
Resident	Capital	40	7	4	2
	City area	12	3	0	0
	Metropolitan	7	0	0	1
	Rural	5	0	0	0
	Immigrant	1	0	0	0
Education	Post Graduate	17	3	0	1
	Undergraduate	41	5	4	1
	Hsc and Below	6	1	0	1
	Not mentioned	1	1	0	0

Table 3.1: Survey and interview participation by category

6 participants whose educational qualification was HSC and below. Undergraduates had a significant presence in the case of interviews (5).

Chapter 4

Interview and Survey Findings

In the following sections, we will get into the specifics of the survey results by analyzing the data we have collected. We reviewed previous works relevant to our study topic and intended audience to make the survey questionnaire easy to comprehend and informative. We created the questionnaire in both English and Bengali because our target audience is from Bangladesh and because we must remember that our target audience is older and may have varying levels of digital literacy. The answers we got were both interesting and thought-provoking, showing how different the participants' points of view were. We tried to understand their psychology of interest in the topics they often search for or like to watch. We also put them as examples in the questionnaire, such as sports, religion, politics, and worldwide current affairs. We divided the questionnaire into three parts. i. Participant Demographics ii. YouTube Usage iii. Examples and opinions.

4.1 Findings from Statistical Analysis

Throughout this research, we have conducted qualitative and quantitative surveys/interviews to understand our target audience's perspective when encountering misinformation on YouTube. We have conducted an online survey to gather information and opinions from 65 participants. This analysis revealed that 47.3% of respondents spend at least three hours on the platform. When asked what is the primary reason for using YouTube, 61.9% of them chose entertainment, as seen in figure 4.1, followed by educational videos at 42.9%. Most participants(54%) actually search for specific topics when they browse YouTube. Meanwhile, 50.8% just view whatever the home page suggests. Also, 74.6% are subscribed to some channels, which shows us that they like regular updates and videos from that specific channel. When asked how they interact with the content posted on YouTube, 55.6% answered that they press 'Like' on the videos, and 23.8% actually save the video for viewing later.

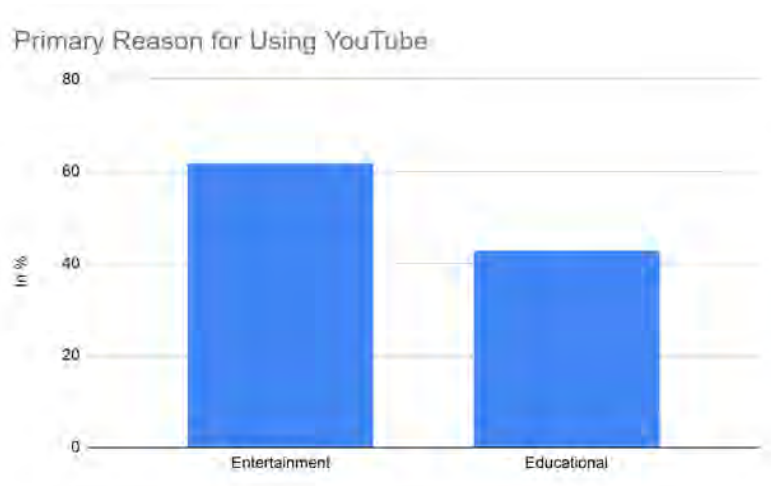


Figure 4.1: YouTube usage percentage

After collecting this general usage information, we presented them with some examples of Fake/misleading/ suspicious titles/videos containing suspicious thumbnails. The participants' reactions to these contents varied according to age and education level. Those aged between 40-50 were less likely, and 51-60 were more likely to be misled by any emotional or similar thumbnails and titles, as seen in figure 4.2.

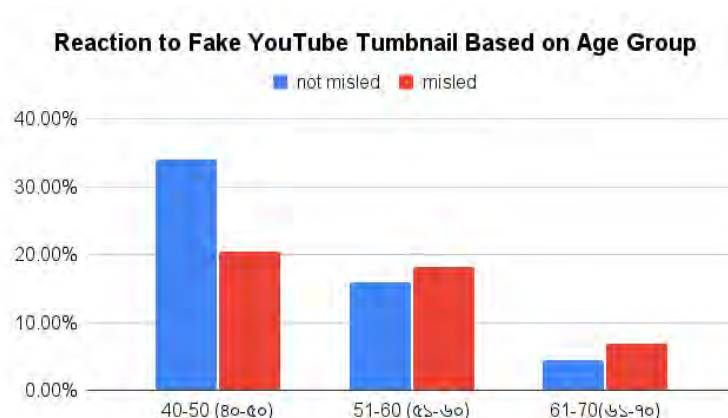


Figure 4.2: Age-Based Reactions to Fake YouTube Thumbnails

Education level showed a high level of influence during this part, as those with undergraduate or postgraduate degrees showed higher skepticism towards misleading videos, they were more aware when it came to believing misleading content, as highlighted in the figure 4.3

Participants were also questioned if they would share this type of video with others or not. Surprisingly, about 39% of them answered that they would share this video to spread awareness regarding misinformation being spread across YouTube. On the other hand, 19.5% said that they might share it in general without any specific intention. A handful of participants actually responded by saying that they would

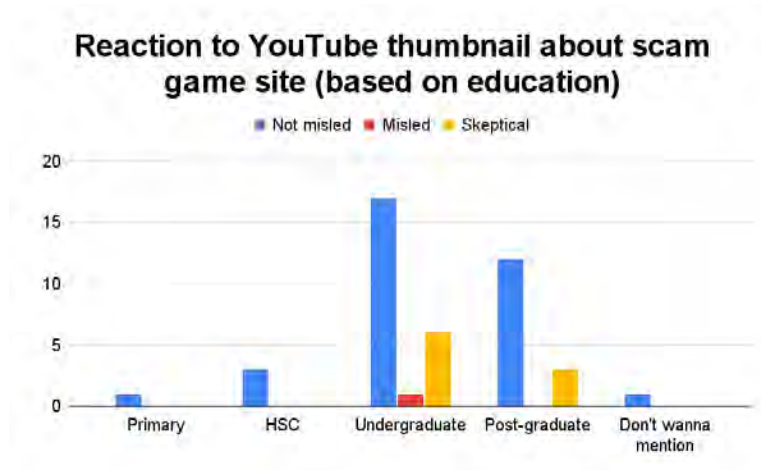


Figure 4.3: Educational Impact on Reactions to Scam YouTube Thumbnails

share this information with their friends, family, and colleagues, either to warn them or because they think that content might be of interest to them.

4.2 Findings from Thematic Analysis

A thorough analysis of our interviews with 17 participants showed that the participants face several issues almost daily while browsing YouTube. This analysis reveals some common misconceptions and trust issues regarding trusting a publisher and users' difficulties navigating YouTube content. We have conducted a mix of semi-structured interviews, focus group discussions, and in-person interviews (door-to-door interviews), which covered a wide range of participants from various age groups and educational backgrounds. The interview questionnaires were designed in such a way that allowed the participants to express themselves candidly. Focus group discussions focused more on how each other viewed their way of regulating through YouTube. Most participants described their struggles when it came to distinguishing between misleading content, primarily when a clickbait title or thumbnail was used. The interview results revealed that participants with higher educational backgrounds and experience using social media for a longer time could identify misleading content. Others remained highly skeptical, mainly when the video was emotionally charged. This varied methodology helped us gather extensive insights into how the older generation of social media users interact with YouTube, which also raises a concern about the necessity for better awareness and protective measures.

4.2.1 Misconceptions Regarding YouTube Thumbnail vs Content

Throughout the interview, we observed that participants often needed help differentiating between videos with eye-catching thumbnails, as shown in Figure 4.4 and actual videos with credibility. Misleading thumbnails, mainly designed to attract as



Figure 4.4: Fake Thumbnail example

many views as possible, were engaging and interesting to our interview participants. Several interviewees admitted clicking on videos based solely on the thumbnail and titles without paying attention to the channel or user it was published from. But they also said that watching such videos with eye-catching thumbnails and titles often turned out to be fake, unrelated to the title or thumbnail, or fabricated after they invested some time into that video. Most of our interviewees answered that they've been watching YouTube videos for at least three years or more, but still, these misconceptions were seen among them.

4.2.2 Lack of Knowledge of News/Articles/Published Dates

During our qualitative interviews, when the participants were shown a video containing a celebrity or politician, they tended to believe that news must be from a trusted source as that person is influential and nothing fabricated can be published about them. We made an interesting observation during the interview, which was the lack of attention they provided to the publication date of a video they were watching, which is an important form factor when it comes to watching fabricated news. Most participants admitted ignoring irrelevant information such as the published date, where it was published, etc, which led them to believe the fabricated videos which contained misinformation. This lack of awareness caused confusion and led the participants to provide opinions based on old or irrelevant data. The interviewees also admitted that they simply watched a video because it was on their YouTube homepage or suggestions page without verifying the source or date.

4.2.3 Trust Issues Rooted from Bad Experiences

We have observed that most of the interviewees developed some trust issues from previous experiences or incidents. Participants from all age groups shared stories or incidents from their earlier experiences of being scammed or someone they know faced difficulties. These experiences varied from being scammed by links posted or shown in videos to believing and sharing a fabricated story or video that they

later got to know was fake. Many of our interviewees mentioned that these previous experiences, where they were deceived by misinformation or fabricated information, made them skeptical about most videos they watch on YouTube. This skepticism although sometimes can be helpful, also made it very difficult for participants to differentiate between videos with good credibility and videos containing misinformation, as demonstrated in 4.5, where an authentic news video thumbnail is shown.



Figure 4.5: Authentic news.pdf

4.2.4 Shyness and Discomfort

The biggest problem we encountered throughout our interview was getting the participants to agree to an interview. After a brief discussion, some participants revealed their discomfort regarding the interview process as it's based on their personal experience and consumption of YouTube. We even had to interrupt an interview mid-way because a participant was unwilling to answer the questions and because the participant believed that their data might be misused, even though we convinced them several times and ensured that their response would not be shared anywhere without their consent. When asked about some of our participants' unwillingness to share their opinions, they expressed that they were easily deceived by fake content, which embarrassed them. Several participants also mentioned feeling embarrassed or hesitant to ask for help evaluating any information they received from any video on YouTube. This shyness and discomfort led to acceptance of content from any source, leaving the participants more vulnerable to manipulation through fabricated videos.

4.2.5 Limited Knowledge of Social Media Manipulation Techniques

Our participants showed limited knowledge when asked what they understood about the term 'Fake content.' Almost all of the participants had some idea regarding what is fake content, at least to their understanding. Their knowledge was minimal and they lacked technical and various tactics such as deepfake, rumors, satire posts, and

edited or cropped videos. When told to differentiate between authentic and cut or cropped videos, they expressed difficulties. These cropped videos are made for the sole purpose of promoting a specific agenda or narrative. Unfortunately, most of the time, these videos are perceived as authentic. Most participants noted that they had seen several videos that they thought were edited or cropped from another video or source. Still, they were unsure of how to verify their authenticity. This lack of technical knowledge and understanding increased their susceptibility to believing and sharing these videos.

4.2.6 Combating The Spread of Misinformation

Our final part of the interview was collecting their feedback on how the platform looks and how it can be improved. Also, they were questioned on how they're combating misinformation they're encountering, what steps they have taken, and whether they know what to do to stop this from happening in the future. Unfortunately, most participants had no idea how to tackle this widespread issue of misinformation. When asked if they reported any video they found irrelevant, most said they did, whereas some had no idea how to report a video or where the report option is located in the interface. Some participants showed interest when asked who should be responsible; when it came to this topic, most of the participants agreed that taking provocative measures by YouTube's authority should be the top priority. They also highlighted the need for verifying videos before they reach the audience, emphasizing the role of these tech companies in combating misinformation.

Chapter 5

System Implementation

To combat the issue of fake videos deceiving old people, we propose an extension. The purpose of the extension is to create a barrier so that we can warn the users if the video they are watching is fake or not. So, by using our extension, they are getting clear guidance and indication about fake content. On the other hand, after knowing the concept of the extension, older users will feel more confident than before while browsing YouTube. Since they are getting labeling for each video, they can make wise decisions by themselves in choosing whether to believe the video content.

5.1 Design Proposal A Real-time Extension of “Bangla Shield.”

We concluded our findings that people in their old age are less knowledgeable in the field of fake and misinformation on YouTube. Due to that, they are more likely to fall victim to fake content and misinformation and get scammed out of money in the worst-case scenario. To protect them we have come up with the design of Bangla Shield, an extension to detect fake content on YouTube. For detecting fake and misleading YouTube videos using our extension, we needed to train an NLP with the best fit for our purpose with a dataset. As our primary focus is on Bangladeshi users, we created our dataset with Bangla, Bangla sentences written in English or Bangla mixed with English, and a few English comments in the videos we gathered them from. The videos are posted on YouTube in the Bangla language.

Methodology

1. Survey and Interviews: To understand the behavior and opinions of old people regarding their user experience on YouTube, we use some methods to collect data from them to implement tools. We conducted quantitative surveys, semi-structured interviews, proxy interviews, and group discussions. In the survey, we provide a set of questions and based on their answer we collect a good amount of data to compare with all responses. We got some valuable information and tried to find a solution that could solve their problem of using YouTube. On the other hand, we used some sampling methods for the interviews. Such as Snowball sampling, convenience sampling, and purposive sampling. We also arranged some proxy interviews so that we could arrange more data for our research. We used several methods so that we could easily



Figure 5.2: Extension Icon

analyze and find out the lacking points and the main target point to work on it.

2. Create an extension: We created an extension where it checks the link of a YouTube video and its send it to server for check. After that server checking process done for that links and it does a calculation and send the calculation data to extension part again. Depending on the calculation there will be a pop up window that where the user can see whether the video link is real or suspicious. By this extension user can choose the videos which are real and can avoid those misleading content. Also if they wants there will also a option where its shows the details about the percentage of real/fake, highlights the top comments by opening a new tab, as demonstrated in figure 5.1, which illustrates the extension's workflow.

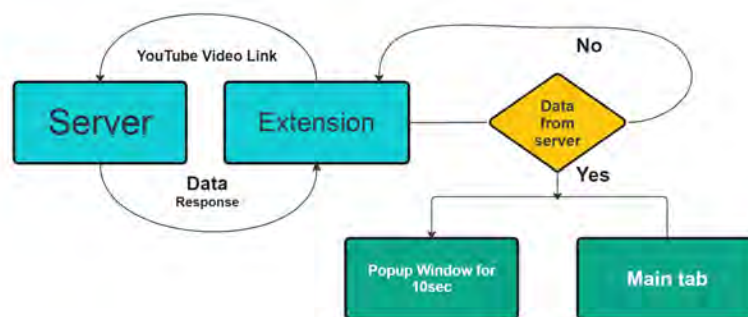


Figure 5.1: Extension Workflow

5.2 Architecture Of Our Developed Extension

Bangla Shield extension extracts URL from the window of Chrome and checks if the URL is for a YouTube site. If the URL is for YouTube the extension sends this link to the server. It waits for a response and if it gets a response it shows a pop-up window with a message and updates the main window. When the user clicks the extension icon, it opens a new tab and shows the details of the video. After that, it waits for a new URL request by the user. In figure 5.3 we get an overview, of how our extension works.

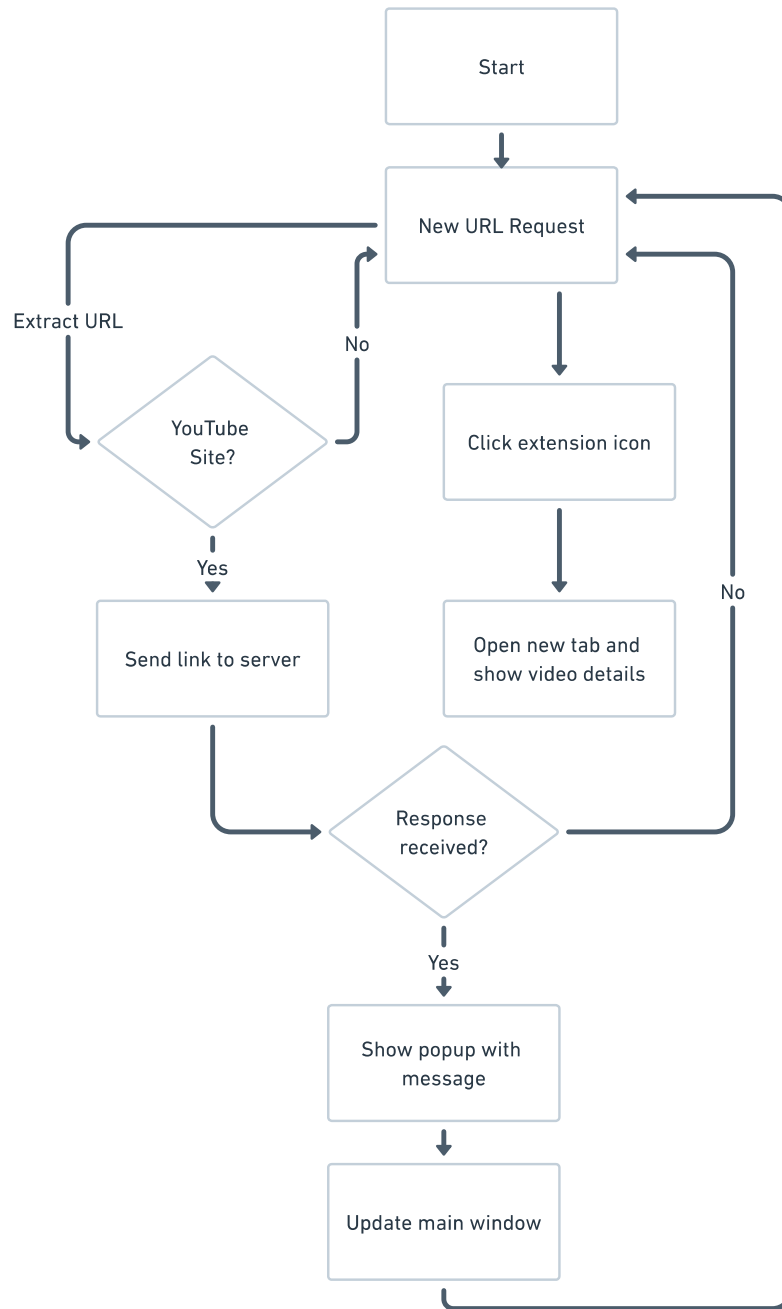


Figure 5.3: Extension Flowchart

5.2.1 Extension Frontend

The front end is the visual presentation of our extension. We tried to keep our design and user interaction as simple as possible. We also tried to give users warning messages with simple text and colored backgrounds. Our front end has two parts.

- Pop-up Window
- Detail Tab

We used HTML, CSS, Javascript, and JSON files to create our front end.

5.2.2 Pop-up Window

Our pop-up window is the primary function of our extension. It gives a warning to users about the YouTube video they are watching. The pop-up window shows four messages in accordance with the likeliness of a YouTube video being fake. Green indicates the lowest probability bracket, yellow indicates the low-mid probability bracket, orange indicates the high-mid probability bracket and red indicates the high probability bracket. The warning is divided into four parts that are, “The video is probably rounded Percentage% fake”, “The video is probably rounded Percentage% False”, “The video is rounded Percentage% suspicious, Be Alert” and “The video has rounded Percentage% probability of being Fake, Avoid it”. The warnings are divided into different percentage ranges, as shown in table 5.4

Bracket	Percentage Range	Color Name	Message	
Low Probability	0 to 7	Green	ভিডিওটি {rounded Percentage}% মিথ্যা বলে মনে হচ্ছে।	The video is probably {rounded Percentage}% fake
Low-mid Probability	8 to 18	Yellow	ভিডিওটি {rounded Percentage}% ভুয়া হতে পারে।	The video is probably {rounded Percentage}% False
High-mid probability	19 to 37	Orange	ভিডিওটি {rounded Percentage}% সন্দেহজনক সচেতন থাকুন।	The video is {rounded Percentage}% suspicious. Be Alert
High probability	38 to 100	Red	ভিডিওটি {rounded Percentage}% সম্ভাবনা মিথ্যা হওয়ার এড়িয়ে চলুন।	The video has {rounded Percentage}% probability of being Fake. Avoid it

Figure 5.4: Color Segment Table

Percentage range calculation:

We tested 20 fake and 20 Non-fake videos. From there we gathered their fake match percentages to calculate the Percentage range, as illustrated in figure 5.5.

Now,

Non-fake Match Percentages = [38, 16, 15, 9.68, 6.67, 22, 9, 12, 29.03, 7, 21, 13, 14, 15, 13, 19, 18, 8.57, 12, 17.14]

Fake Match Percentages = [8, 83.33, 14.63, 46, 42, 75, 81.81, 71.43, 23, 23, 61.81, 61.54, 50, 46, 42.11, 52.94, 26.09, 23, 21, 19]

Range for low probability bracket = 0 to min(Fake Match Percentages)-1
 = 0 to 8 - 1
 = 0 to 7

The range for high probability bracket = max(Non-fake Match Percentages) to 100
 = 38 to 100

Percentages between low and high brackets: [14, 14.63, 15, 15, 16, 17, 17.14, 18, 19, 19, 21, 21, 22, 23, 23, 23, 26.09, 29.03, 38]

Thus Median: 19.0

So Low-mid Probability = min(Fake Match Percentages) to Median - 1
= 8 to 19 - 1
= 8 to 18

And High-mid probability = Median to max(Non-fake Match Percentages) - 1
= 19 to 38 - 1
= 19 to 37



Figure 5.5: Warning Range: Percentage

When the pop-up window gets the fake match percentage, it checks which warning is applicable and shows the user the warning window for 10 seconds. After that period ends, it removes itself.

```
1  popUp = createPopUp()
2  setupStyleForpopUp()
3
4  // Conditional logic based on Fake Percentage
5  if Fake Percentage < start of Low-mid Bracket then
6      setStyle(popUp, "backgroundColor", "#28a745") // Green
7      setStyle(popUp, "color", "white")
8      TEXT = "ভিডিওটি " + Fake Percentage + "% মিথ্যা বলে মনে হচ্ছে।"
9  else if Fake Percentage < start of High-mid Bracket then
10     setStyle(popUp, "backgroundColor", "#ffc107") // Yellow
11     setStyle(popUp, "color", "black")
12     TEXT = "ভিডিওটি " + Fake Percentage + "% ভুল হতে পারে।"
13 else if Fake Percentage < start of High Bracket then
14     setStyle(popUp, "backgroundColor", "#fd7e14") // Orange
15     setStyle(popUp, "color", "black")
16     TEXT = "ভিডিওটি " + Fake Percentage + "% সন্দেহজনক, সচেতন থাকুন।"
17 else
18     setStyle(popUp, "backgroundColor", "#dc3545") // Red
19     setStyle(popUp, "color", "white")
20     TEXT = "ভিডিওটি " + Fake Percentage + "% সম্ভাবনামিথ্যা হওয়ার, এড়িয়ে চলুন।"
21 end if
22
23 // Display the popUp with the text
24 displaypopUp(popUp, TEXT)
25
26 // Wait 10 seconds and remove the popUp
27 wait(10000)
28 remove(popUp)
```

Figure 5.6: Pseudo-code of Pop-up Window

Here, Figure 5.6 is the pseudo-code for our implemented logic.



Figure 5.7: Samples of Pop-up Window

Figure 5.7 is a demonstration of what the four pop-up windows look like. Here we can see that the pop-up window is showing a message with green, yellow, orange, and red backgrounds corresponding to the match percentage it received from the server,

5.2.3 Detail Tab

The detail tab is another function of the extension. When the user clicks the extension icon it opens a new tab in the Chrome window. It is there to show the user the detailed report of the YouTube video that it gets from the server reply. The information includes a Fake comments percentage with a color background similar to the pop-up window, video title, comments checked, fake comments, likes, dislikes, processing time, and a maximum of 15 sample comments flagged as Fake for users to make their judgment.

Here Figure 5.8 is a view of the detail tab on how users would see the details of the video received from the server. It displays Fake comments percentage with a color background similar to the pop-up window, video title, comments checked, fake comments, likes, dislikes, processing time, and a maximum of 15 sample comments flagged as Fake for users.

Fake Comments: 37.93%

Analysis Details

Video Title: লাইভে এসে কেঁদে কেঁদে সাকিব আল হাসানকে নিয়ে একি বললেন নাকিসা কামাল | Nafisa Kamal | Sakib Al Hasan

Comments Checked: 58

Fake Comments: 22

Fake Percentage: 37.93%

Likes: 1114

Dislikes: 26

Popup Time: 16/10/2024, 14:27:33

Processing Time: 4.595 seconds

Sample Fake Comments:

1. কুজা ও চায় চিত হইয়া শুইতো!!! সঠিক উচ্চারণ করতে পারে না,আবার ভিডিও বানায় !!!!! মিথ্যা তথ্যে ভরা, সেই কথা আর নাই বললাম !!!!!!!
2. View Jonno manuah koto kisu kore
3. Gojob
4. Apnara Jara ashob vuwa khobor Nia ashen tader porinam same e Hobe.
5. bal er news
jor kore kaoke dhore rakha jaina
purus 4 biyer jonno Allah bolechen ei jonnoi
6. এসব ভূয়া নিউজ
7. Dhuro asob fake kahani,, hudai gujob chorai
8. এটা ফেইক ভিডিও।
শিশির ও খুব ভালো করে জানে সাকিব খুব ভালো স্বামী
9. ফালত থাম
10. কোন ভিডিও বা কোন প্রমাণ কোন ভয়েস কারো কোন কথা ছাড়াই এতগুলো কথা বললেন কোন প্রমাণ দেখাতে পারবেন
11. What a report !!
12. Garbage news!! Please stop!
13. সব ফালতু খবর
14. আমার কাছে সব মিথ্যা, ভুল এবং বানানো মনে হচ্ছে।
15. Amr mone hoy ai fulto news na kora, Togo kono kam niy to ke ar korbe,

Figure 5.8: Sample Detail Tab

5.3 Extension Backend

For our extension to run and provide the functions we wanted, backend support was needed. For that purpose, we created a server using Python that would help us with the calculations and comment detection. Our server listens for requests made by the extension. It receives the YouTube video link sent by the extension as a request and extracts the video ID from that. It collects the video title, like count and 100 comments from YouTube using YouTube API. It also extracts dislike count using ReturnYouTubeDislike API. Afterward, it cleans the YouTube comments which are top 100 by relevance. Then the comments are sent to a pre-trained mBERT model for classification in a loop. The model classifies the comments and returns classification and match confidence. If the classification is fake it increases fake_count by 1. After that, if match confidence is above 0.90, it adds the comment to the fake_comment list. If match confidence exceeds 0.70, it adds the comment to the backup_fake_comment list. After that, if fake_comment does not contain 15 comments, it takes the rest from sorted backup_fake_comments. Figure 5.9 is visual of our server-side workings

It also calculates the fake percentage.

$$Fake_Percentage = (Fake_Count/Comments_Checked)100\% \quad (5.1)$$

Finally, it sends all the data to the extension, completes the request, and then listens for new requests.

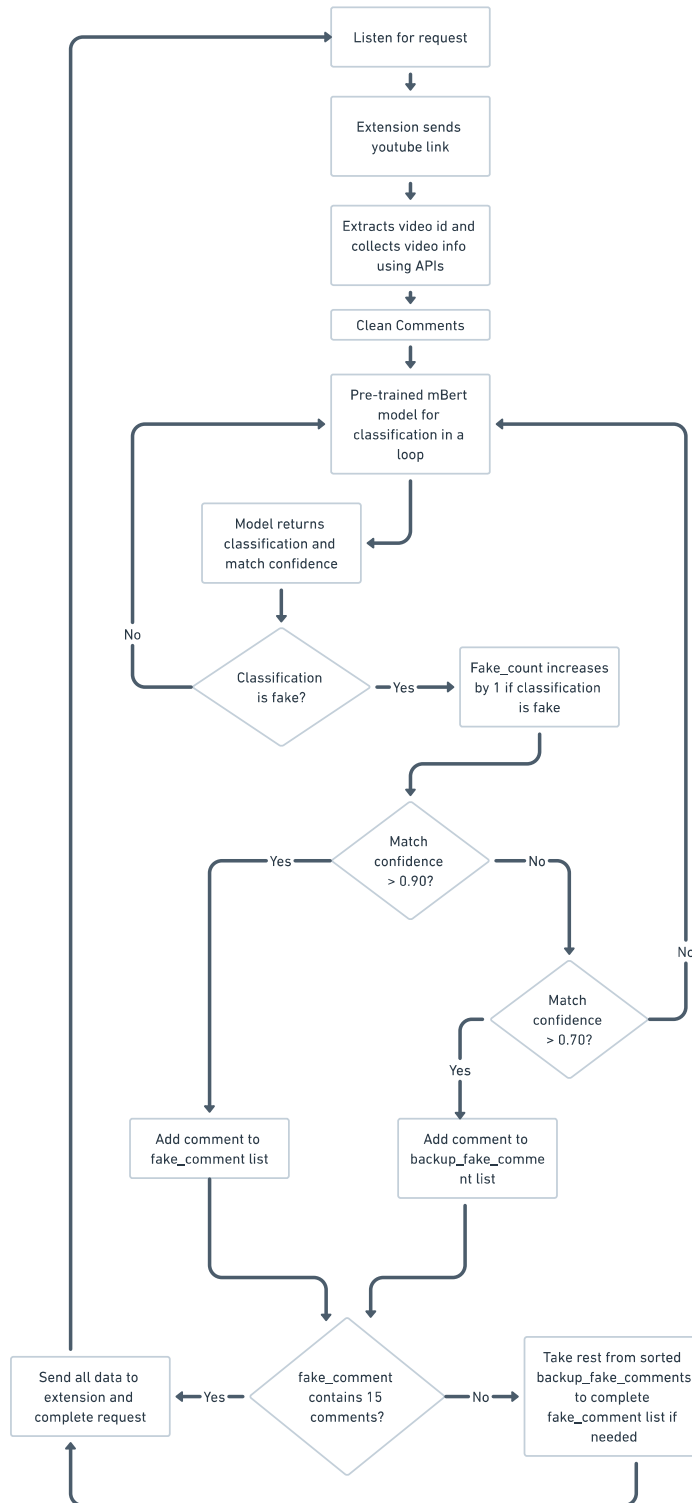


Figure 5.9: Server Flowchart

5.4 Dataset Collection

In order to build our fake video detection extension, we have collected data and made a dataset from YouTube videos which will help us to justify whether our extension is working perfectly for all sorts of videos or not. We collected our dataset based on video titles, views, and comments. In order to build our fake video detection extension, we have collected data and made a dataset from YouTube videos which is very important to verify the effectiveness of our extension. It will also help us to justify whether our extension is working perfectly for all sorts of videos or not. We collected our dataset based on video titles, views, and comments. After gathering the videos and developing the extension it will show the result by popping up the labeling of Fake or Non-fake. Here, by mentioning Fake or Non-fake we mean that Fake videos: false, misleading, or inaccurate information that tries to misrepresent and scam users, Non-fake videos: are videos that provide truth and accurate information about the post that is reliable for the user. Figure 5.10 shows how we collected our dataset. A detailed method of our data extraction is provided below:

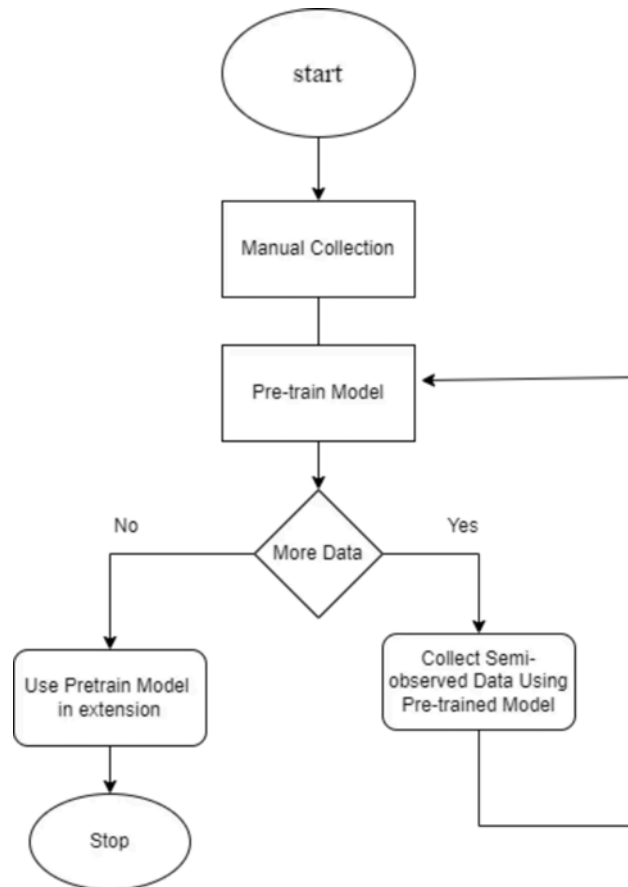


Figure 5.10: Data Collection Flowchart

5.4.1 Data Labels

Our dataset has two columns labeled "Comment" and "Label." The "Comment" column contains all the dataset's comments. The "Label" column has two classes, "Fake" and "Non-fake," which classify the comments.

Fake: Comments that claim the YouTube video is fake or misleading

Non-fake: Positive, neutral or discussion about the video in the comment section of the YouTube video

5.4.2 Initial Dataset

At the start of the journey of creating a custom dataset, we manually gathered more than 600 comments. We sorted them again to find recurring patterned comments and took 411 as our first dataset to pre-train and test. Of the 411 comments, 271 claimed the videos were fake and were labeled as Fake, and 140 were neutral and optimistic about the video and were labeled as Non-fake.

5.4.3 Pre-train Model

Using our first dataset we trained mBERT to get a pre-trained model and used that to gather semi-observed data.

5.4.4 Semi-observed Dataset Collection

Using this pre-trained model, we extracted a maximum of 500 comments per YouTube video and had them labeled by the model. Then, we manually went through them and checked the labeling. We also removed empty and corrupt rows from the dataset. When we find a wrong assumption made by the model, we edit that and put the right label. We went through 4 iterations of semi-observed data collection to make our dataset. Figure 5.11 is the pseudo-code for our semi-observed data collector. As shown here, We labeled a comment as Fake here when comment match confidence is above 60% and if classified as Fake. If not, then comment match confidence is above 30% and is classified as Non-fake. If neither condition matches for the comment we label that as Uncertain to judge that ourselves. After that, the collector stored the comments it classified in an Excel file for us to recheck the labeling.

```

1  function fetchComments(video_id)
2      comments = []
3      loop until comments are 500
4          append comments from response
5          if no next_page_token then break
6      end loop
7      return comments
8
9  function cleanComment(comment)
10     return cleaned comment
11
12 function classifyComment(comment)
13     // Prediction Non-fake = 0, Fake = 1
14     if comment confidence >= 0.6 and prediction = 1 then
15         return "Fake"
16     else if comment confidence >= 0.3 and prediction = 0 then
17         return "Non-fake"
18     else
19         return "Uncertain"
20
21 function extractVideoID(video_link)
22     patterns = ["pattern1", "pattern2", "pattern3", "pattern4"]
23     loop through patterns
24         if match pattern then return video_id
25     return None
26
27 function storeComments(comments, labels)
28     append to excel [comments, labels]
29
30 function classifyYouTubeComments(video_link)
31     video_id = extractVideoID(video_link)
32     if video_id is None then return
33
34     comments = fetchComments(video_id)
35     if comments are empty then return
36
37     fake_count = 0
38     loop through comments
39         label = classifyComment(comment)
40         if label = "Fake" then fake_count = fake_count + 1
41
42     match_percentage = (fake_count / length(comments)) * 100
43     storeComments(comments, labels)
44
45 loop forever
46     video_link = input("Enter link (or 'q' to quit): ")
47     if video_link = "q" then break
48     classifyYouTubeComments(video_link)

```

Figure 5.11: Semi-Observed Data Collector Pseudo-code

5.4.5 Final Dataset

At the end of our data collection, we have collected 11010 comments to train a model to use in our extension. Out of classified 11010 comments, there are 8247 comments are Non-fake and 2730 comments are Fake.

Total Comment	Fake Comment	Non-fake Comment
11010	2730	8247

Table 5.1: Summary of Comments

5.5 Model Training and Testing

To implement our proposed model, we needed to train our dataset. We then examined how ML and NLP models worked and how they could help us to implement our extension.

5.5.1 Natural Language Processing(NLP)

We have tested three NLP models to train our dataset that supports the Bengali language. The tested models that support our requirements are Bangla-Bert, mBERT, and XLM-R. But first, let's start with an idea about how NLP works. Natural Language Processing(NLP) uses theories and technologies to enhance the interpretation between computers and humans, with the goal of enabling computers to understand human language as input and generate appropriate output.[23] [13] These are some of the important levels of NLP:

- Phonology: It involves the study of organizing sound within a system.
- Morphology: Examines the forming of words from basic but meaningful units.
- Lexical Analysis: Investigates and identifies the structure of words within a paragraph.
- Syntax Analysis: Involves studying the structure of sentences to identify relationships among words through systemic arrangement.
- Semantics: This determines whether the given word possesses a suitable meaning.
- Discourse: It involves examining the significance found within the current, prior, and following sentences.
- Pragmatics: Understanding the true significance of a statement within its context. [15]

5.5.2 Machine Learning

Machine Learning(ML) is a part of Artificial Intelligence that aims to develop systems that can learn and make decisions based on data without any external programming. This involves developing algorithms and techniques that can use traditional programming methods, allowing the system to “learn” from provided data and improve performance over time.[23]

There are two primary methods of learning:

- Supervised Learning: This method involves training the model with labeled input and output data, which enables it to provide predictions based on the feedback received. It is used to create decision trees and neural networks. Training stops once the target accuracy has been reached.

- Unsupervised Learning: In this learning model, the system learns from provided data without external guidance, which helps in clustering and association tasks. Popular algorithms for these clustering and association tasks include the Apriori Algorithm.[23] [11]

Algorithms:

- K-Nearest Neighbor(KNN): KNN is one of the most simple methods used for classification. It involves the nearest training data points to the test point and assigns labels based on majority voting. [27]

$$d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (5.2)$$

- Support Vector Machine: SVM is a tool used for both linear and non-linear classification. It divides the data into two possible clauses using a hyperplane, categorizing it as a non-probabilistic binary linear classifier.[4]
- The logistic regression(LR) model is a widely used statistical method for binary classification in supervised machine learning. It anticipates whether an event belongs to one of two categories. The model estimates the probability of a given input that belongs to a specific class, using a sigmoid function to map input data to a probability range of 0 to 1.[65]
- Decision Tree: Decision Trees are a simple yet effective method for characterizing smaller datasets. The computational unpredictability increases significantly with the number of measurements of the information. Extensive datasets often result in convoluted trees, which require a considerable amount of memory for storage. [69]

We have tested our dataset on logistic regression and SVM models as well with the other NLP models.

5.5.3 Graphs

Classification report: A classification report evaluates the performance of a classification model through various metrics such as Accuracy, Precision, Recall(Sensitivity), F1-score, etc. These metrics come from the confusion matrix, which contains true positive(TP), true negative(TN), false positive(FP), and false negative(FN) values. Each of these metrics serves a different purpose in assessing the model's efficiency in classifying data:

- Accuracy: The ratio of correct predictions to the total number of predictions.

$$\text{Accuracy} = (\text{True Positives} + \text{True Negatives}) / \text{Total Predictions}$$

$$= (TP + TN) / (TP + TN + FP + FN)$$
- Precision: It is defined as the ratio of true positive predictions to the total number of predicted positives, calculated as $(TP / (TP + FP))$. This metric measures the accuracy of positive predictions.

- Recall(Sensitivity): This refers to the model’s ability to accurately identify all relevant instances, calculated as $(TP/(TP+FN))$.
- F1-Score: It’s the balanced mean of Precision and Recall, which provides a balance between the two metrics, especially in the context of imbalanced datasets.[58]

It is calculated as $F1 = 2 * (\text{precision} * \text{recall}) / (\text{precision} + \text{recall})$
 $F1 = 2TP / (2TP + FP + FN)$

ROC: A receiver operating characteristic(ROC) curve is a graphical representation tool that is used to observe the performance of a diagnostic test. The relationship between sensitivity(true positive rate) and 1-specificity(false positive rate) is illustrated across various threshold values of a test, resulting in continuous outcomes. This curve helps determine how efficiently a test discriminates between different states, such as the presence or absence of a condition.[18]

Confusion Matrix: Confusion matrix is a 2D array used to determine the performance of a classification method by summarising the predicted classification compared to the actual classifications. It provides counts of correct and incorrect predictions, as well as the calculation of metrics such as accuracy, sensitivity(true positive rate), and sensitivity(true negative rate). In binary classification, the confusion matrix is represented as a 2x2 table; for multiple categories, it is expanded accordingly. A confusion matrix is useful because it can provide a detailed breakdown of a model’s classification performance, which enables it for a better evaluation of accuracy, precision, recall, and error types, which eventually helps in fine-tuning the models and also helps understand their strengths and weaknesses.[42]

Learning Curve: The learning curve concept is that when the quantity of units produced is doubled, the direct labor hours required decrease at a constant rate(e.g., 90%, 80%, etc.). This occurs following the formula $Y = KX^n$, where Y represents the labor hours for the head unit, K is the hours for the first unit, X is the cumulative number of units, and n is the learning index. This indicates that enhanced efficiency results come from both labor learning and organizational learning processes. [1]

Accuracy vs. Epoch: The accuracy vs. epoch graph illustrates the correlation between model accuracy and the number of training epochs. Accuracy measured tells the model’s ability to predict correct outcomes, while epochs indicate the number of times the learning algorithm processes the entire training dataset. This graph is calculated by tracking the model’s performance over each epoch, which displays the accuracy of both training and validation datasets. It’s essential for model training as it helps visualize the model’s learning progress, indicating instances of under-fitting and over-fitting and whether further training could improve or harm the model’s performance.[19]

5.5.4 Bangla-BERT Model

Bangla-BERT is a monolingual BERT model designed explicitly for Bangla Language. The model is based on the Transformer architecture, commonly used for a range of NLP tasks. Unlike the multilingual BERT(mBERT), which uses weights for various languages, Bangla-BERT is specifically designed and pre-trained on an extensive dataset which is dedicated to the Bangla language, utilizing about 40GB of text from different sources. It helps Bangla-BERT to be more effective in tasks

related to the Bangla language, including sentiment analysis, binary classification, and named entity recognition.

Bangla-BERT’s pre-training contains two significant tasks:

1. Masked Language Modeling(MLM): This involves the random masking of tokens within a sequence and predicting the masked tokens.
2. Next Sentence Prediction(NSP): This predicts the relationship between two provided sentences.

This model uses Byte-pair Encoding(BPE) for tokenization and embedding, generating numerical representations of tokens within the text. The scaled dot-product attention formula is a fundamental component of Bangla-BERT; it helps the model determine the significance of various tokens related to each other.

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V \quad (5.3)$$

Here,

- Q, K, and V are the query, key, and value matrices.
- dk Represents the dimensionality of the key vectors.

Bangla-BERT outperforms any other Bangla language models that include mBERT and other non-contextual models, like Bangla fasttext and word2vec, in downstream NLP tasks. Its binary classification, sentiment analysis, and named entity recognition performance increase significantly, making it the best Bangla-NLP model.[54] We’ve trained Bangla-BERT using our dataset and generated a classification report, accuracy vs epoch graph, confusion matrix, and ROC curve. The results are:

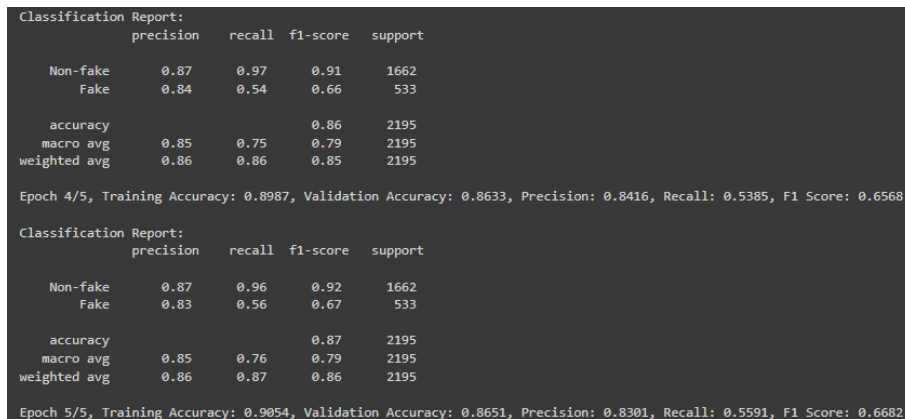


Figure 5.12: Bangla-BERT Training Results

In this output Figure 5.12, we can see the detailed accuracy metrics of the Bangla-BERT’s last two Epoch results. It shows individual results of the Fake and Non-fake classes. Epoch 4 and Epoch 5 both have similar accuracy of 90%. Epoch 4 has a precision of 84%, recall of 54%, and f1 score of 66% for Fake classification. On the other hand, Epoch 5 has a precision of 83%, recall of 56%, and f1 score of 67% for Fake classification.

So, Epoch 5 is better than Epoch 4 by a small margin.

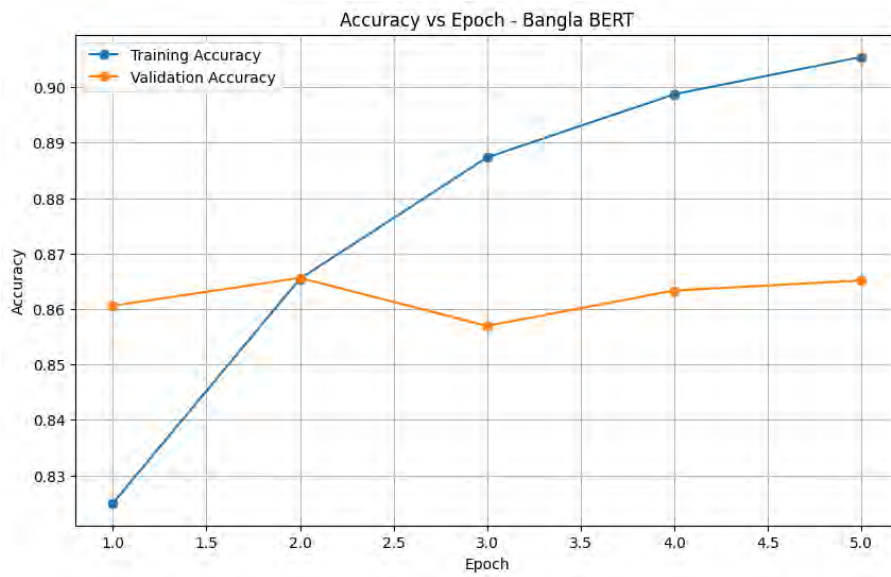


Figure 5.13: Bangla-BERT Accuracy Vs Epoch

This Bangla-BERT accuracy vs Epoch Graph in Figure 5.13 shows training accuracy and validation accuracy of Bangla-BERT over 5 Epoches. We see that accuracy increases in a downturn curve as the increment in accuracy decreases from 83% in Epoch 1 to a little over 90% in Epoch 5. We also notice the model generalizes data well but struggles to maintain that as we see a drop in validation accuracy in Epoch 3.

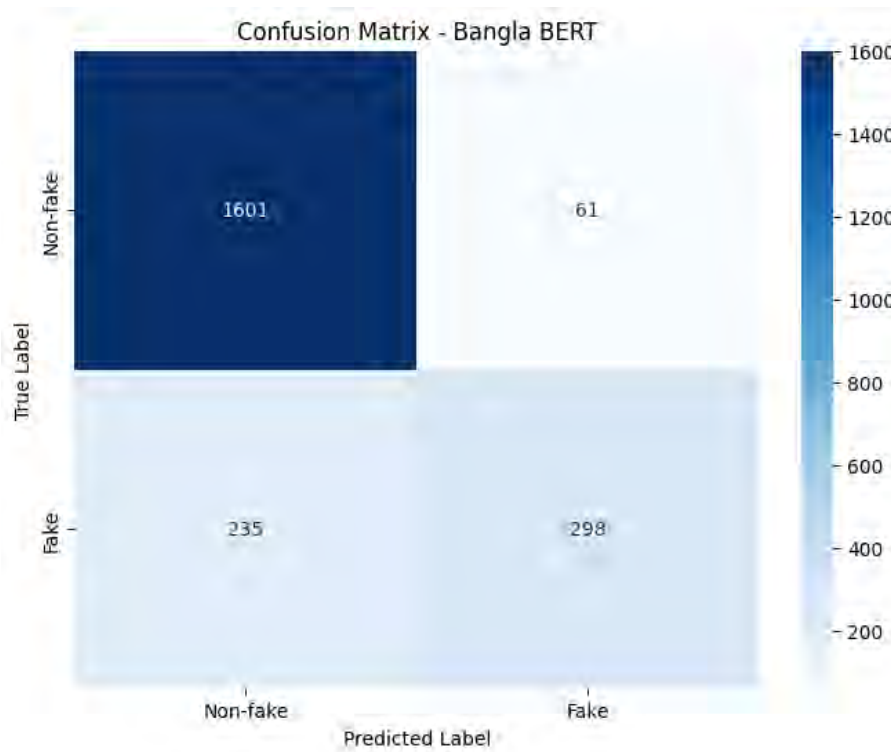


Figure 5.14: Bangla-BERT Confusion Matrix

According to the confusion matrix shown in Figure 5.14, Bangla-BERT classifies Non-fake comments very well, only classifying false negatives of 61 comments out of the 1662 Non-fake instances of data classification and guessing the rest of the 1601 comments correctly. It is a different story in the Fake classification as we see that out of the 533 tested comments, it predicted only 298 of them correctly as fake and gave false positives in 235 other predictions which is a huge drop in accuracy in Fake classification.

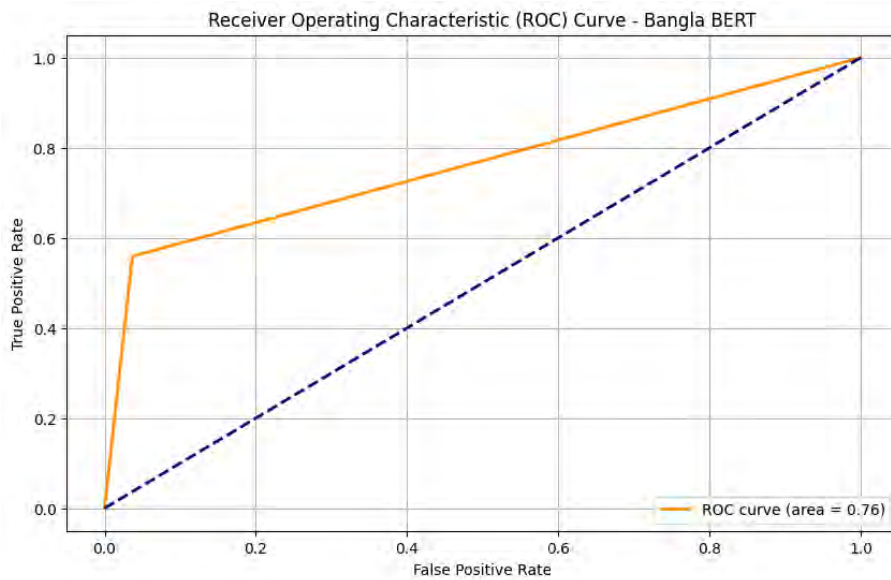


Figure 5.15: Bangla-BERT ROC Curve

In this ROC curve illustrated by Figure 5.15, the orange line indicates how effectively Bangla-BERT can differentiate the Fake class and Non-fake Class.

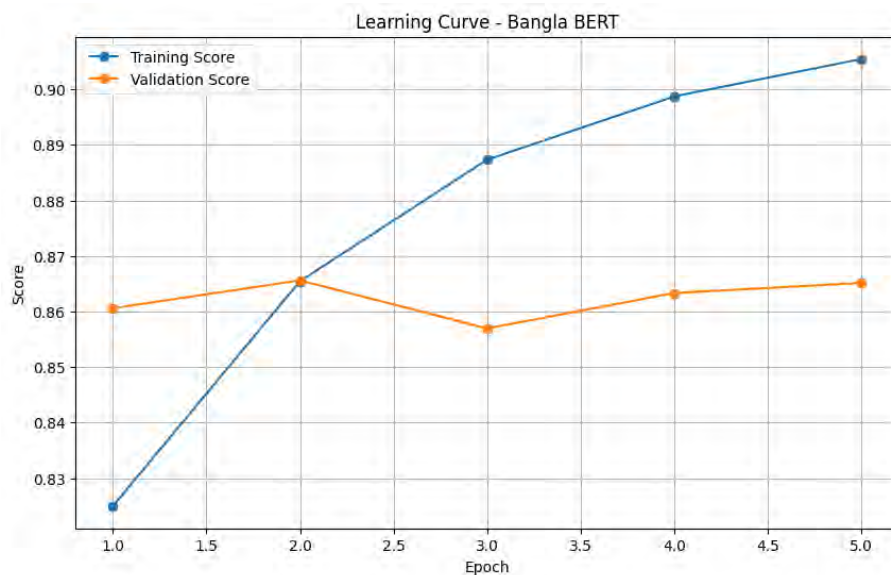


Figure 5.16: Bangla-BERT Learning Curve

Figure 5.16, shows the AOC is 76%. Thus we can tell, that it can correctly classify a comment 76% of the time and the other 24% times it gives false classification.

5.5.5 mBERT Model

Multilingual BERT(mBERT), developed by Devlic et al.(2019), is a language model that has been pre-trained on monolingual corpora across 104 languages. This is effective when it comes to zero-shot-cross-lingual model transfer, utilizing task-specific annotations in one language to fine-tune the model to conduct evaluation in a different language. mBERT generates multilingual representations across various scripts and typologically similar languages, though it still has several limitations in accommodating specific language pairs.[32]

The working structure of mBERT consists of training a 12-layer transformer model on combined monolingual Wikipedia corpora from 104 languages using a shared word-piece vocabulary. mBERT lacks the implementation of particular language indicators or methods to ensure translation-equivalent representations. This approach allows cross-lingual generalization by effectively capturing multilingual representations, which allows for efficient zero-shot model transfer across various languages, even those with different scrips, although it shows optimal performance with topologically similar languages.[32]

We have generated a classification report, accuracy vs. epoch graph, confusion matrix ROC curve, and learning curve by training mBERT using our dataset.

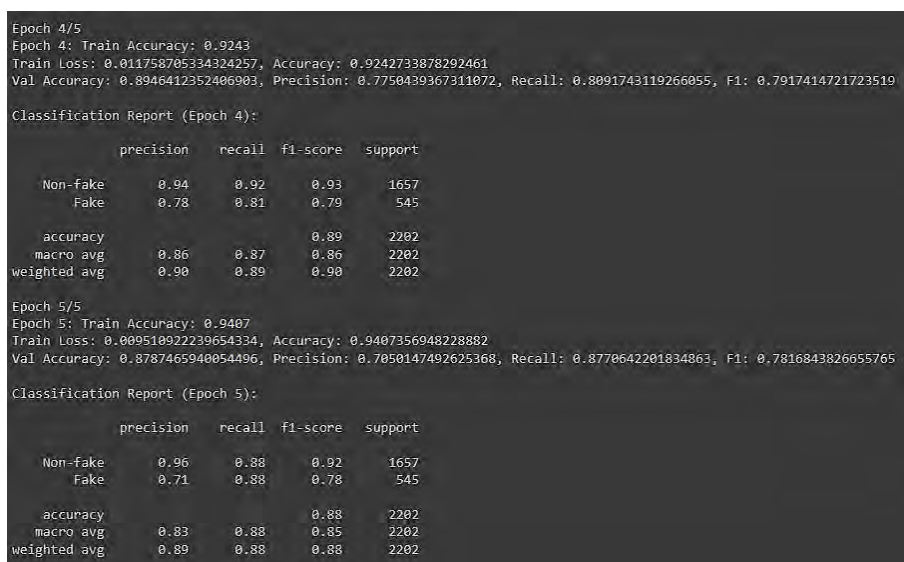


Figure 5.17: mBERT Training Results

In this Figure 5.17 of mBERT training results, we can see the last two Epoch results of the mBERT. It shows detailed accuracy metrics of Fake and Non-fake classes. Epoch 4 has an accuracy of 92.43% which is lower than Epoch 5 accuracy of 94.07%. Epoch 4 has better precision(78%) and worse recall(81%) compared to Epoch 5 precision (71%) and recall(88%) for fake class detection. In f1-score Epoch 4 is ahead with 79% compared to Epoch 5 with 78%. Overall, Epoch 4 is balanced in terms of precision, recall, and f1-score at the cost of little accuracy for detecting fake class compared to Epoch 5.

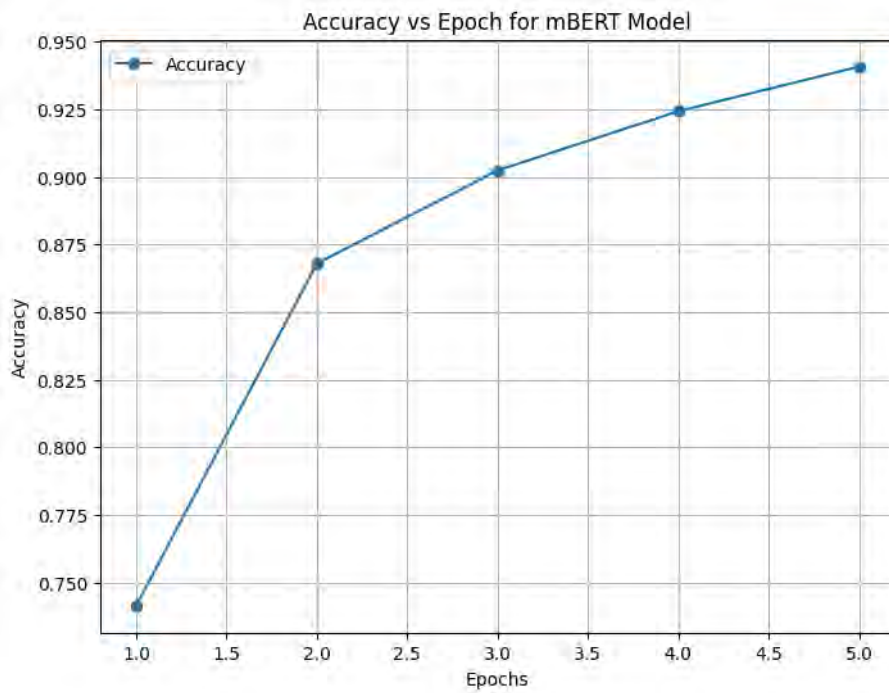


Figure 5.18: mBERT Accuracy vs Epoch

Figure 5.18 shows a graph where the mBERT model’s accuracy is increasing with every Epoch. We see a sharp increase in accuracy from Epoch 1 accuracy of 75% to Epoch 2 accuracy of 87%. We also can tell from the graph that mBERT achieved 90% accuracy after the 3rd Epoch and had the maximum accuracy of 94% at Epoch 5. We also observe a decrease in accuracy improvement as Epoch number increases as it is nearing 100% and the upward trend slows down.

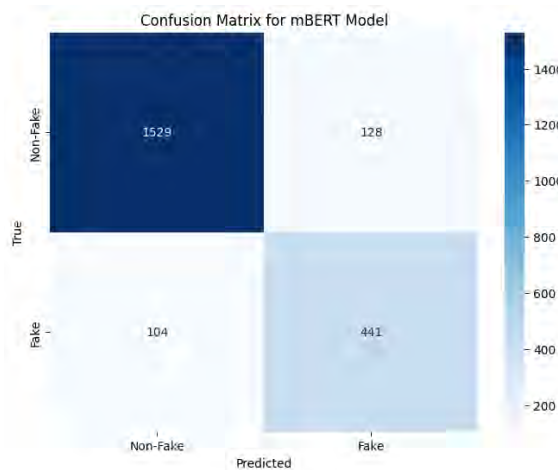


Figure 5.19: mBERT Confusion Matrix

In the confusion matrix in Figure 5.19, we observe that of the 1637 Non-fake instances of data classification, the model correctly predicted 1529 of them and gave false positives 128 times. In the case of Fake predictions we see that out of the 545 tested comments, it predicted 441 of them correctly as fake and made mistakes in 104 other instances.

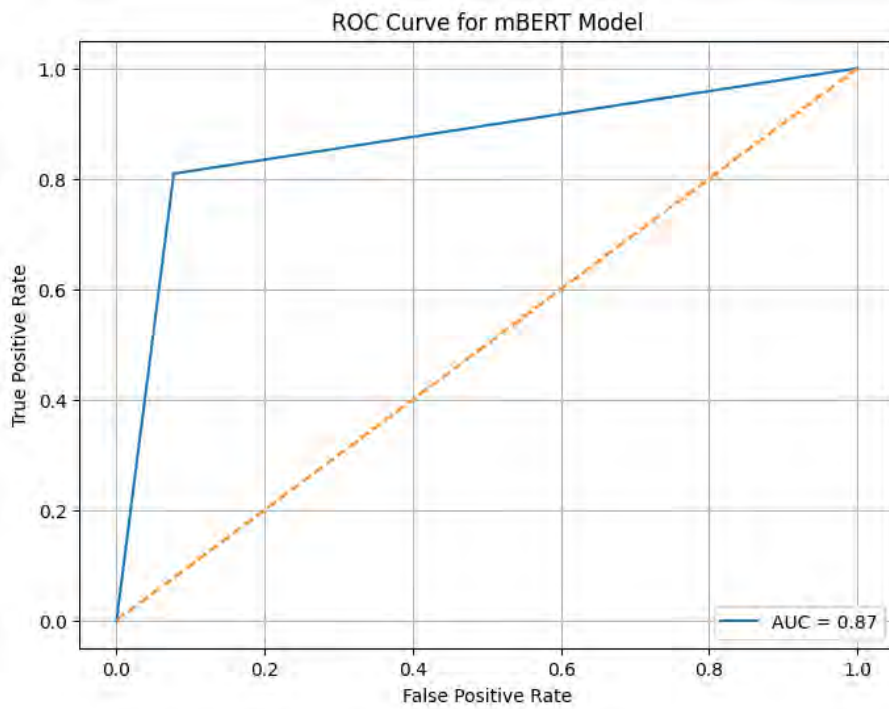


Figure 5.20: mBERT ROC Curve

In this ROC curve represented in Figure 5.20, the blue line shows how effectively mBERT can differentiate the Fake class and the Non-fake class. Here the AOC is 87%. This means it can 87% of the time correctly classifies a comment's class.

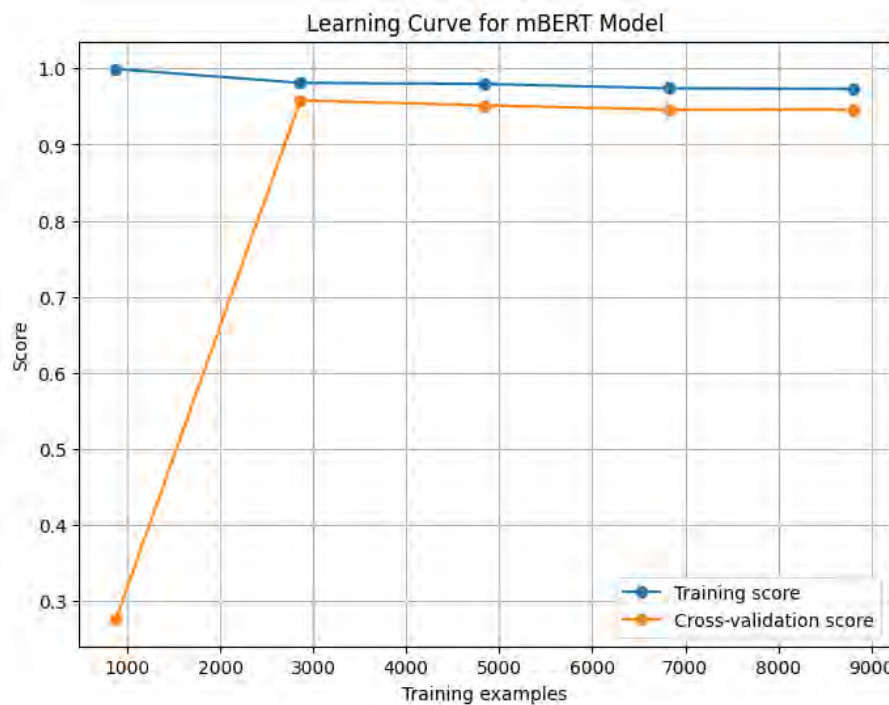


Figure 5.21: mBERT Learning Curve

The learning curve of mBERT shown in Figure 5.21 illustrates an improvement in accuracy as the number of training data increases. According to the chart, we can

conclude that after nearly 3000 data accuracy stabilizes and the model finds it hard to generalize the dataset under 3000 data. But after 3000 training data, mBERT generalizes well and gives a consistent result.

5.5.6 XLM-R Model

XLM-R is a transformer-based multilingual masked language model that has been trained on data from 100 languages using CommonCrawl data. The model utilizes the Transformer architecture and implements a masked language modeling(MLM) objective to predict tokens that have been randomly masked in the text. Some key features of this model include:

- Training Data: XLM-R uses an extensive corpus from CommonCrawl, covering 100 languages, which improves performance for low-resource language in comparison to mBERT
- Vocabulary and Tokenization: Uses SentencePiece for tokenization, featuring a vocabulary of 250,000 tokens, which allows effective management of multilingual text and code-switching.
- There are two variants- XLM-R Base, which consists of 12 layers and 270M parameters, and XLM-R Large, which contains 24 layers and 550M parameters.

XLM-R’s performance has achieved an average accuracy increase of 14.6% on XNLI, a 13% average F1 improvement on MLQA, and a 2.4% increase on NER relative to m-BERT, making it practical for cross-lingual tasks while maintaining pre-language performance.[29]

By training XLMR with our dataset, we generated a classification report and graph showing accuracy vs epoch, confusion matrix, and ROC curve.

```

Classification Report:
      precision    recall  f1-score   support

 Non-fake      0.94      0.90      0.92      1684
 Fake          0.72      0.81      0.76       512

 accuracy      0.88      0.88      0.88      2196
 macro avg     0.83      0.86      0.84      2196
 weighted avg  0.89      0.88      0.88      2196

 Predictions distribution: [1620 576]

Classification Report:
      precision    recall  f1-score   support

 Non-fake      0.94      0.90      0.92      1684
 Fake          0.72      0.82      0.76       512

 accuracy      0.88      0.88      0.88      2196
 macro avg     0.83      0.86      0.84      2196
 weighted avg  0.89      0.88      0.89      2196

 Predictions distribution: [1615 581]

```

Figure 5.22: XLM-R Training Results

Here in Figure 5.22, we can see the last two Epoch results of the XLM-r model. It shows detailed accuracy metrics of the Fake and Non-fake classes. Epoch 4 and Epoch 5 have an accuracy of 88% precision of 72% and f1-score of 76% for fake class detection. In recall, Epoch 5 is ahead with 82% compared to Epoch 4 with 81%. Overall Epoch 4 and Epoch 5 give more or less the same results in both Fake and Non-fake classification.

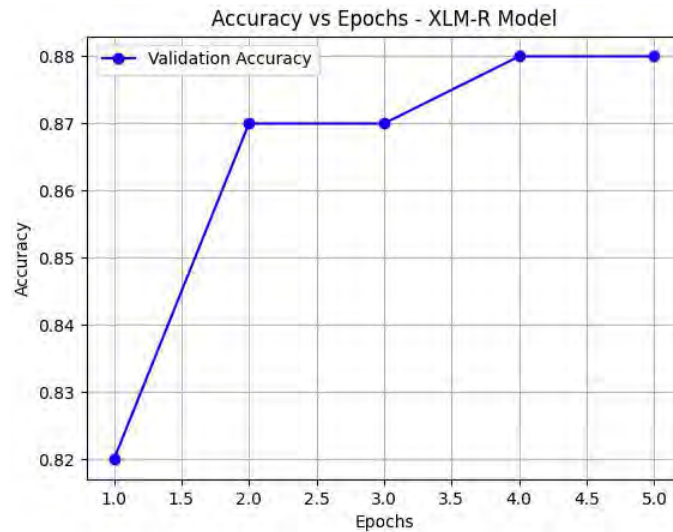


Figure 5.23: XLM-R Accuracy vs Epoch

In this graph depicted in Figure 5.23, we observe that the XLM-R model's accuracy is increasing with every two Epochs. We see a sharp increase in accuracy from Epoch 1 accuracy of 82% to Epoch 2 accuracy of 87%. We also can tell from the graph that XLM-R achieved 88% accuracy after the 4th Epoch, and stabilized and maintained the same accuracy after Epoch 5.

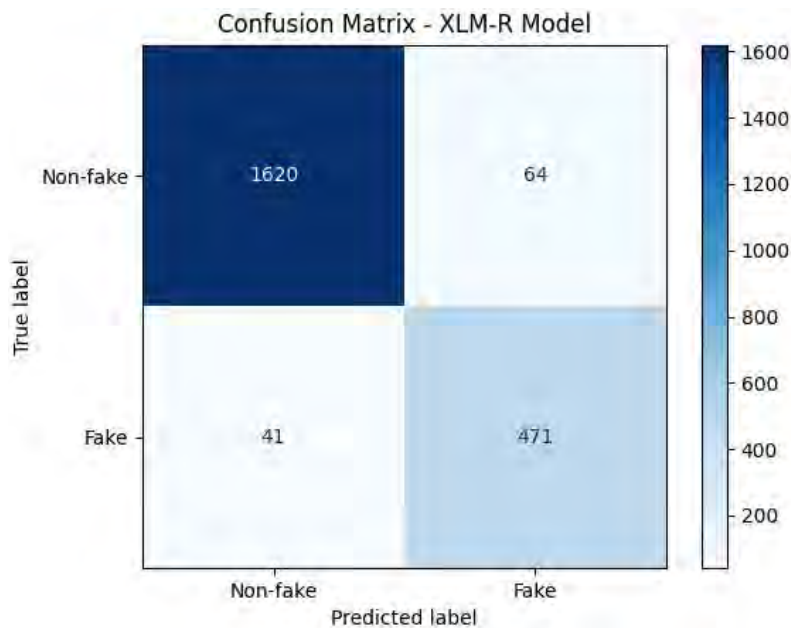


Figure 5.24: XLM-R Confusion Matrix

Figure 5.24 shows a XLM-R confusion matrix where XLM-R classifies Non-fake comments very well, only classifying false negatives of 64 comments out of the 1684 Non-fake instances of data classification and guessing the rest of the 1620 comments correctly. It is a similar story for the Fake classification as we see that out of the 512 tested comments, it predicted only 471 of them correctly as Fake and gave false positives in only 41 other predictions, which is a very good accuracy in Fake classification.

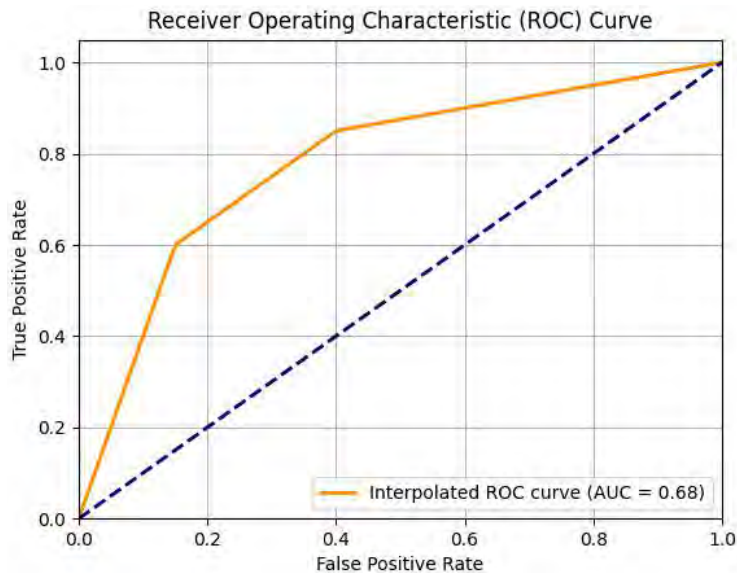


Figure 5.25: XLM-R ROC Curve

The ROC curve of XLM-R portrayed in Figure 5.25 shows with the blue line that it can 68% accurately predict the Fake and Non-fake class.

5.5.7 Logistic Regression Model

The logistic regression model is used as a regression method when the outcome variable is binary or dichotomous. The model delineates the association between a dependent variable and a collection of independent variables.

The logistic regression model is expressed in the following form:

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} \quad (5.4)$$

Here, $\pi(x)$ denotes the conditional probability of the outcome being 1, given the predictor variable x .

The logistic regression model uses the logistic function to ensure that the predicted probabilities are always in the range of 0 to 1.[7]

Using our dataset, we trained a logistic regression model, generating a classification report, accuracy vs epoch graph, confusion matrix, ROC curve, and learning curve.

```

PS C:\Users\zoniae\OneDrive\Documents\Thesis Defense\Thesis models\Thesis models\Logistic Regression> python logistic_regression.py
Label distribution in the dataset:
Label
0      8245
1      2727
Name: count, dtype: int64
Total samples: 10972

Training set class distribution:
Label
0      6596
1      2181
Name: count, dtype: int64

Test set class distribution:
Label
0      1649
1       546
Name: count, dtype: int64
Overall Accuracy: 0.8264
Precision (Fake): 0.6391
Recall (Fake): 0.6941
F1-Score (Fake): 0.6655

Classification Report:

```

	precision	recall	f1-score	support
Non-fake	0.90	0.87	0.88	1649
Fake	0.64	0.69	0.67	546
accuracy			0.83	2195
macro avg	0.77	0.78	0.77	2195
weighted avg	0.83	0.83	0.83	2195

Figure 5.26: Logistic Regression Train Result

From the output snippet shown in Figure 5.26, we see that after training over a dataset of nearly 11000 comments, Logistic regression has an accuracy of 82.6%. Though it has overall good scores classifying precision of 90%, recall of 87%, and f1 score of 88%, it has poor results of precision 64%, recall 69%, and f1-score of 67% for detecting Fake class.

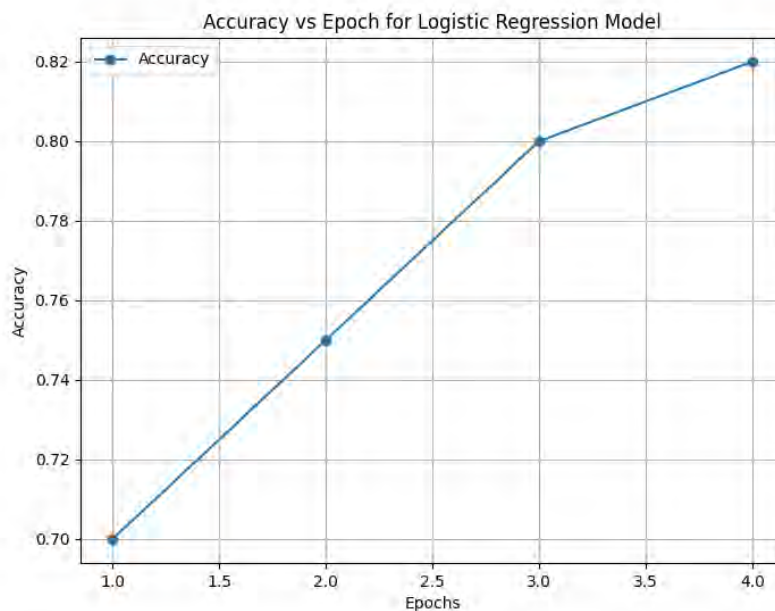


Figure 5.27: Logistic Regression Accuracy vs Epoch

The graph represented in the Figure 5.27, shows that the Logistic Regression model's accuracy is increasing with every Epoch. We observe an increase in accuracy from Epoch 1 accuracy of 70% to Epoch 3 accuracy of 82% at Epoch 4. The accuracy

increase will further drop for future Epochs as accuracy gets closer to 100% or comes to a stalemate in an accuracy improvement.

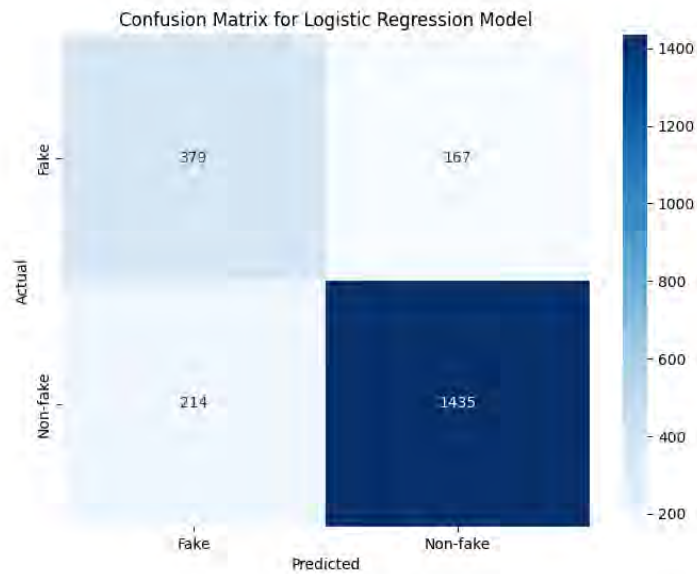


Figure 5.28: Logistic Regression Confusion Matrix

In this confusion matrix seen in Figure 5.28, we can observe how well Logistic Regression Identifies the Fake class and the Non-fake class. Like other models it also classifies Non-fake comments well, classifying false negatives of 214 comments out of the 1649 Non-fake comments, and predicting the rest of the 1435 comments correctly. Like Bangla-BERT It has poor Fake classification results as we see that out of the 546 comments tested, it classified 379 of them correctly as fake and gave false positives in other predictions which is a good Percentage drop of accuracy in Fake classification.

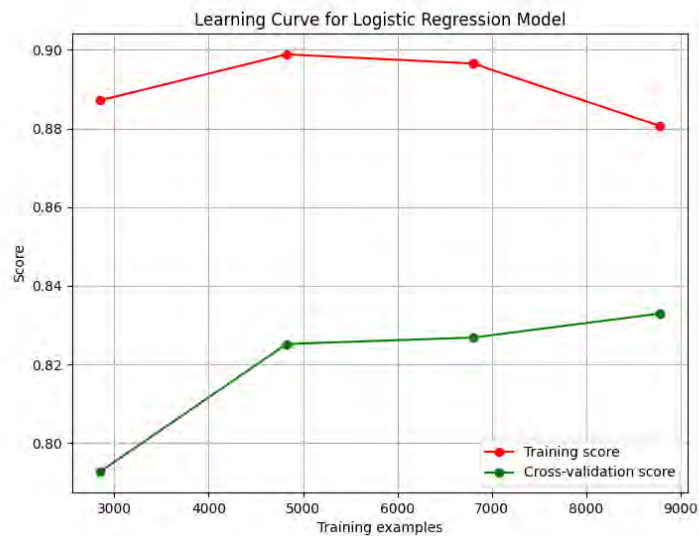


Figure 5.29: Logistic Regression Learning Curve

The learning curve of Logistic Regression portrayed in Figure 5.29 shows the improvement of accuracy as the number of training data increases. According to the chart, we can conclude that after nearly 5000 data accuracy gets maximum value but starts to decline and it declines sharply from 7000 comments to 9000 comments and the model finds it hard to generalize the dataset under 5000 data. But after 5000 training data, it generalizes well and increases the generalization from 7000 comments onward.

5.5.8 SVM Model

A Support Vector Machine(SVM) is a machine learning model based on statistical learning theory that is primarily used for classification and regression-related tasks. The process involves mapping input data into a high-dimensional feature space through kernel functions and constructing an optimal hyperplane to distinguish the data points. SVM seeks to minimize empirical risk while maintaining model complexity to achieve good generalization.

For linearly separable data, the hyperplane can be found by solving an optimization problem in order to maximize the margin between the classes. The optimization uses a cost function:

$$F(W) = \frac{1}{2}W^TW \tag{5.5}$$

subject to linear constraint. In non-linear scenarios, data is transformed into a higher-dimensional space through kernel functions such as polynomial, radial basis function(RBF), or perceptron kernels, while maintaining the application of the same optimization principles applied in this transformed space.[5]

We trained SVM on our dataset and produced a classification report, accuracy vs epoch graph, confusion matrix, ROC curve, and learning curve.

```
PS C:\Users\zoniae\OneDrive\Documents\Thesis Defense\Thesis models\Thesis models\Svm> python svm.py
Overall Accuracy: 0.8465391621129326
Precision (Fake): 0.6746506986027944
Recall (Fake): 0.66015625
F1-Score (Fake): 0.667324777887463

Classification Report:

```

	precision	recall	f1-score	support
Non-fake	0.90	0.90	0.90	1684
Fake	0.67	0.66	0.67	512
accuracy			0.85	2196
macro avg	0.79	0.78	0.78	2196
weighted avg	0.85	0.85	0.85	2196

Figure 5.30: SVM Training Result

From the output screenshot as reflected in Figure 5.30, we find that after training over our dataset SVM has an accuracy of 84.6%. Though it has overall good scores classifying precision of 90%, recall of 90% and f1 score of 90%, like logistic regression, it has poor results in precision of 67%, recall of 66%, and f1-score of 67% for detecting Fake class.

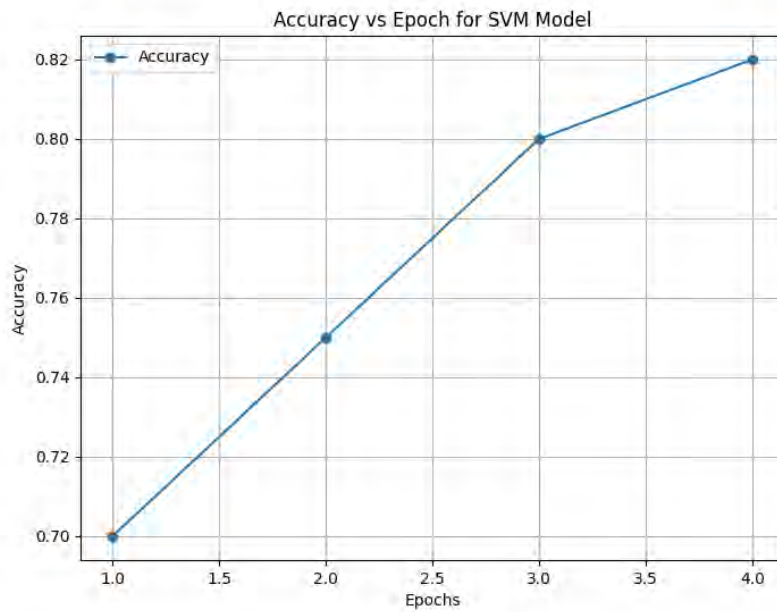


Figure 5.31: SVM Accuracy vs Epoch

The graph in Figure 5.31 shows that the SVM model's accuracy is increasing with every Epoch. We observe an increase in accuracy from Epoch 1 accuracy of 70% to Epoch 3 accuracy of 82% at Epoch 4. The accuracy increase will further drop for future Epochs as accuracy gets closer to 100% or comes to a stalemate in an accuracy improvement.

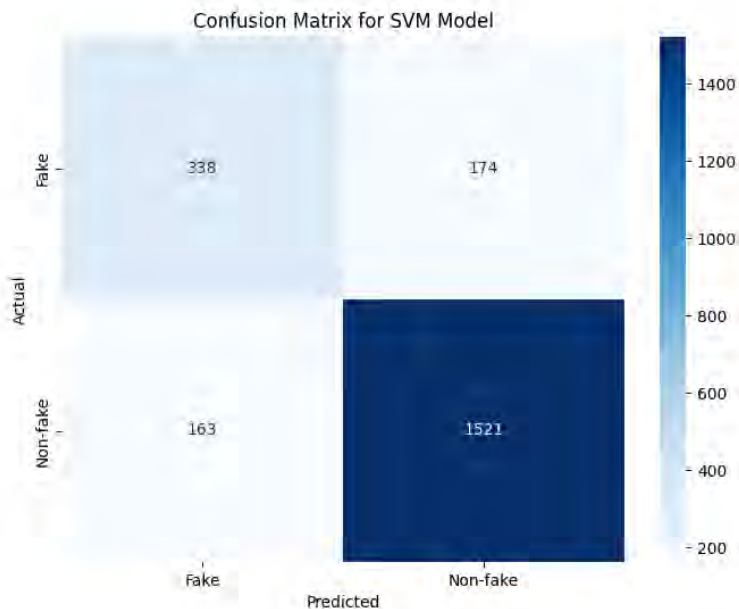


Figure 5.32: SVM Confusion Matrix

In the SVM confusion matrix demonstrated by Figure 5.32, we can see how well SVM performs in differentiating between the Fake and Non-fake classes. It classifies

non-fake comments with good accuracy, like the other models we previously tested. Out of the 1684 non-fake comments dataset, it accurately classified 1521 and for the other 163, it gave false negatives. Out of the 512 Fake comments it tested, it only correctly classified 338 of them as Fake and gave wrong predictions to the other 174. That is a big drop in accuracy. Thus this model is not very accurate at detecting Fake comments.

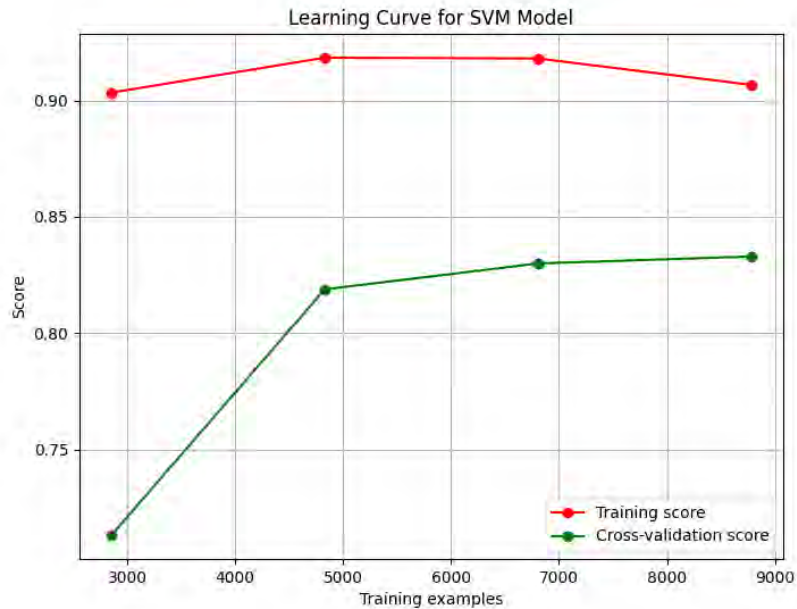


Figure 5.33: SVM Learning Curve

The SVM learning curve reflected in Figure 5.33 shows that with the increase of training data, accuracy also increases. According to the chart, accuracy peaks around 5000 comments but then starts to drop between 7000 and 9000 data points. Also, the model has a hard time generalizing the dataset below 5000 comments. But, after training with 5000 comments, it generalizes well and gains a little more generalization from the 5000 to 7000 comment range and stays at 84% cross-validation.

5.6 Comparison and Findings from Models

To detect fake YouTube videos, we need to classify comments extracted from YouTube videos. To do that, we need a model to classify the comments and detect comments that claim the video is fake. After compiling the training results of all the NLP models and ML mentioned, we made a comparison between them and found the best model that fits our needs.

The table 5.2 is the compilation of the results, we got from training the models with our own dataset. As we see in the table mBERT has the best overall performing model. It has better accuracy (92.4%) second highest precision (78%), highest recall (81%), and f1 score (79%). It has a balance of precision and recall which is missing in other models. Bangla-BERT has the highest precision (83%) but falls behind because of the lowest recall (56%). XLM-R shows a promising percentage in every accuracy metric but it falls short in accuracy (88%) and precision (72%) compared to

Model	Accuracy	Precision	Recall	F1-score
Bangla-BERT	0.87	0.83	0.56	0.67
M-BERT	0.924	0.78	0.81	0.79
XLM-R	0.88	0.72	0.82	0.76
Logistic Regression	0.83	0.64	0.69	0.67
SVM	0.85	0.67	0.66	0.67

Table 5.2: Performance Metrics of Different Models

mBERT. Logistic Regression and SVM both have below 70% in the precision, recall, and f1 scores even though they got accuracy scores of 83% and 85% respectively. 5.34 We concluded through the comparison that mBERT is the best fit for our extension to be used to classify comments and detect fake or misleading YouTube videos.

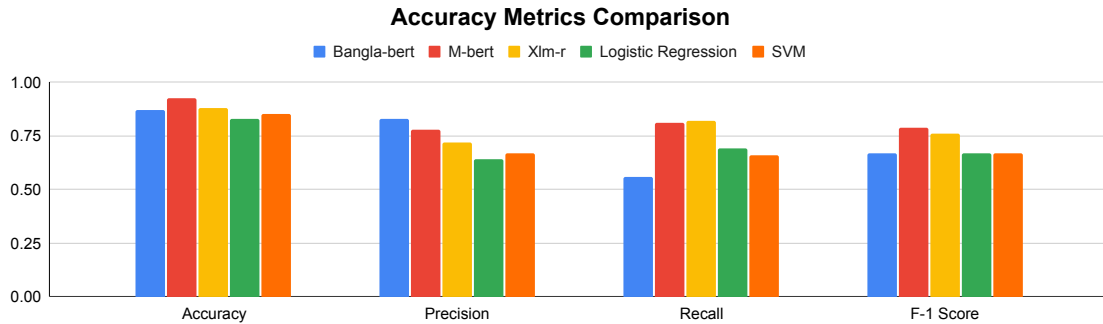


Figure 5.34: Accuracy Metrics Comparison

5.7 Subjective Evaluation

Subjective evaluation refers[2] to the qualitative assessment of the model’s output, particularly in tasks such as machine translation. This evaluation typically involves human judges who assess the quality of translations based on criteria like fluency, adequacy, and overall quality rather than relying solely on quantitative metrics. Subjective evaluations are essential because they provide insights into the model’s performance from a human perspective. They allow for a more nuanced understanding of how well the model captures the meaning and context of the input text. This can help identify strengths and weaknesses that automated metrics may not fully capture. As mentioned beforehand, we trained for different NLP and ML models. During the different testing phase when we fetched files from server.py basically our main backend server from where the extension runs we coded in such a way that it asks for YouTube video link and if the original video has 500 or more comments in it server.py and it fetches 500 comments at max. Initially, when we started collecting almost 600 comments for our dataset we labeled fake and non-fake manually by ourselves, and to avoid bias we tested the dataset manually and then after training initially we gradually progressed with 1700,1610,4000 and 5600 comments. We checked the quality or accuracy of the generated data by our server after 6k dataset comments it started giving better results but when we manually checked the machine-generated labeling we saw discrepancies. We saw comments which the

machine detected as fake but in real it was non-fake. Also, some emojis were detected as fake. We observed these discrepancies and made some changes manually in the dataset and after 6k comments we made the final dataset by performing subjective evaluation to reach our final dataset of 11k+ comments. For this reason, we performed some evaluation in order to check our extension performance towards dedicated fake and non-fake videos picked randomly over YouTube in Bangla language which is 30(15 fake,15 non-fake) in number. Then we used our extension to see what results exactly it gives by color labeling if it matches or not. Surprisingly it gave most of the videos correctly and color labeled almost accurately. But discrepancies hit in video 13 and video 31 where it gave the exact opposite results labeling real non-fake video to orange and fake video to green. That is our one tested observation from the extension's subjective evaluation. Overall, our subjective evaluation has resulted in improving our proposed model. Furthermore, it has helped us identify and improve in terms of detecting Fake and Non-fake comments.

Limitations of "Bangla Shield"

- It is made only for pc version basically browser-base.
- Not yet available in any web store to use as a plugin still a prototype
- Does not give 100% accuracy rate
- If there is any misguided video in which comments are fabricated as non-fake but the original content is fake, it cannot be detected as fake.
- Sometimes under-perform in detecting dedicated non-fake videos
- Can only fetch 100 top comments to calculate the fake percentage
- In the new tab, only 15 fake comments are shown.
- Due to developer limitations extension can process 100 videos per 24 hours

Chapter 6

Extension feedback Survey

As already mentioned in our RQ3, our browser extension “Bangla Shield” opens the door for a new, easy, and better experience against the barriers that our older generation faces in terms of their YouTube usage. Our research question focuses on solving the problem barriers discussed in Chapter 5. During our first course of interviews and surveys, we found that social media users are more into YouTube as the content type is video all over. Its easy-to-use interface allows the user to engage more and more in the content. As a result, users fall for fake content and get manipulated easily. Suppose we want to explore other barriers and try to solve them with our browser extension. In that case, we have been able to detect the fake percentage of the video by our trained model and within our own created dataset especially focusing on our Bangla language-related videos as our older generation is more into exploring Bangla news and related videos on YouTube which we previously found in our interview findings.

Furthermore, this extension is kept simple and in the Bengali language due to the low digital literacy of our participants. To evaluate our solution, we again need feedback to prove the durability of our extension solution. As a result, we designed a questionnaire and did a feedback survey online using Google Forms as the medium. At first, we let our targeted audience know about our extension’s features and then attached screenshots of the extension working in different genres of YouTube, showcasing the popup message and color-labeling(green, yellow, orange, red) according to their fake percentage ratio. After fully demonstrating the extension, we try to understand if the users are satisfied and easily understand the processes we demonstrated. Moving into the next, we try to figure out having a clear idea about our extension if they are willing to use it on their own and recommend it to others. If they liked or found the helpful extension, what is one feature(s) that made our extension eye-catching, and what more can we do to improve our extension or add or modify any feature(s).

6.1 Feedback Survey Questionnaire

To prepare a questionnaire focused on older users, we kept the questionnaire very simplistic, using Google Forms as the medium. We prepared the questionnaire with multiple choice questions, linear scale responses, and a couple of open-ended questions to get their opinion. Therefore, we simply divided the questionnaire into Demographics, Understanding and Interface Clarity of “Bangla Shield”, Feedback,

and suggestions. In the first part, we focused on the demographic information of the participants the usage of YouTube, and some other general questions. For the second part, to make the participants understand our extension interface and working clarity, we added screenshots of different features of our extension we picked a random video from YouTube enabling our extension in the background, and then our first feature, where the video is color labeled showing the fake comment percentage as already illustrated in chapter 5. Then we attached the screenshots of where we clicked on the extension icon and a new tab opened up, retrieving and analyzing different information from the video showing the user 15 fake comments and fake percentages along with some general information. Moving on, we ask them how clear they are about the color labeling concept and the whole concept of the extension. Also asked about their willingness to stop or continue the video even after detecting it fake to understand how clear they are to our extension interface. Lastly, in the Feedback and suggestions part, we asked the participants if they could be cautious using our extension and watching YouTube videos, along with an open-ended question about whether they would recommend our extension to their family and friends and the best feature if they like our extension keeping the option for non-recommendation and disliking as well to avoid bias from the questionnaire.

6.2 Findings from Feedback Survey

We conducted the survey online and kept the form open only for 12 hours, where we got 65 responses from the older users themselves or someone who filled up the form on their behalf as they might need to be more efficient in using Google Forms or filling up. From our previous analysis, we already knew if the solution is given in the Bangla language and kept simple, it would attract the old-aged users for the simplistic outlook of the extension. If we focus on some demographic information, age groups were divided into 41-50,51-60,61-70. Most of the survey participants(76.90%) are men and women from the 41-50 years age group. Among the 65 participants, 47.6% of them are comfortable or very comfortable using browser extensions while using YouTube, whereas 29.2% are in midpoint or confused whether to use the extension comfortably or not as illustrated by 6.1. The most exciting and proven fact came out when participants were asked to tell the type of content they usually see on YouTube. Whopping 72.3% of the older aged users are watching news and politics-related videos on YouTube from where we can easily say that if these news and political videos are made intentionally then old aged users can be misguided easily without detecting the authenticity of the YouTube video.

Even though many people get misguided through YouTube videos they understand the concept of fake videos or misleading content which we saw back in our semi-structured interviews. When asked, 44.6% of the total respondents come across fake YouTube content very often which is an alarming thing.

In the second part of the video, where we portrayed our extension feature as screenshots, 86.10% of the users found the work process of the "Bangla Shield" extension clear or very clear, meaning our extension's simplistic look grabbed their attention and gave them a clear idea.

72.20% of the total respondents found our color labeling system (Green: Safe, Yel-

low: possibly Unsafe, Orange: be careful, Red: Might be fake) useful or very useful to detect fake content along with comment percentage.

Despite the warning system we have put in place, 44.6% of the users still responded that they would continue watching the fake or misguided video even after it is labeled as orange or red having a higher possibility of being fake or misleading.

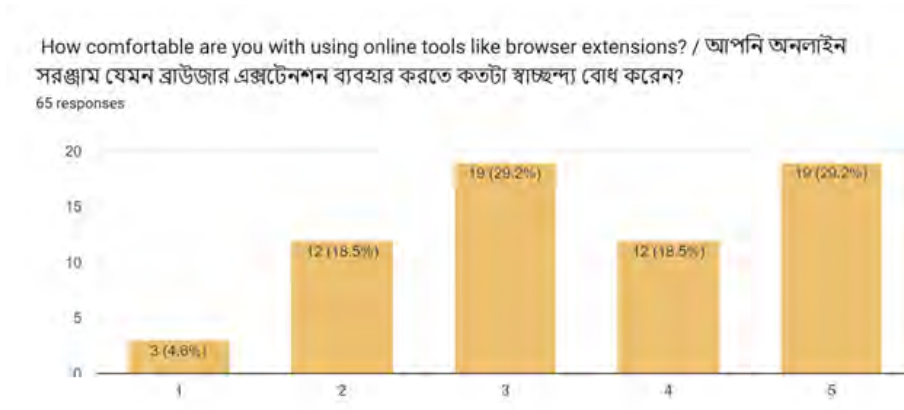


Figure 6.1: Willingness to use Browser Extensions

When we tried to take feedback on your new tab viewing 15 fake comments, fake comment percentage among the top 100 comments, view count, like dislike numbers, 80% of the users found this feature helpful or very helpful.

The bar chart in Figure 6.2 illustrates the helpfulness of our pop-up feature among users, with responses categorized as "Helpful," "Extremely Helpful," "Neutral," and "Not very helpful." The majority of participants found the pop-up "Helpful" (almost 40%), followed by "Extremely Helpful" (13%). A smaller portion of respondents were "Neutral" (almost 10%), and only a few found it "Not very helpful" (less than 5%).

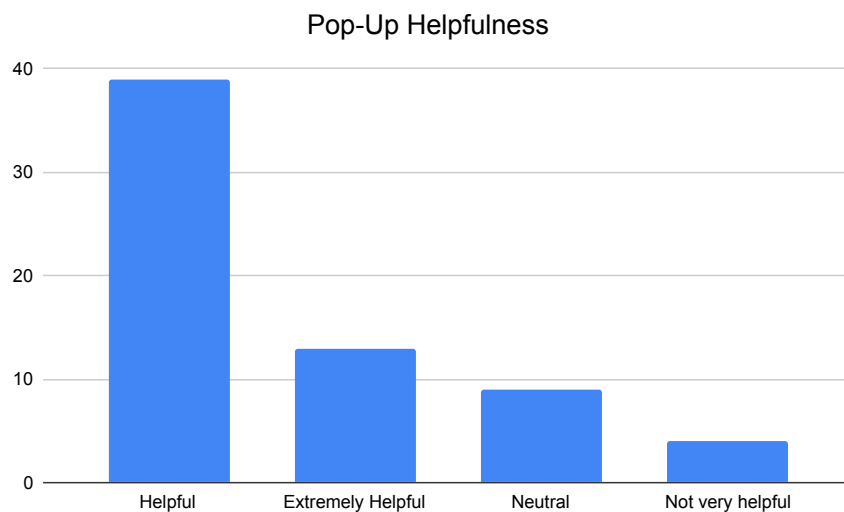


Figure 6.2: Pop-Up Helpfulness

The bar chart in Figure 6.3 shows participants' willingness to detect fake YouTube videos. Most of our participants answered "Yes," which indicates a strong interest in detecting fake videos. A smaller portion of respondents were either "Not Sure" or answered "No", showing less interest in cooperating with such efforts.

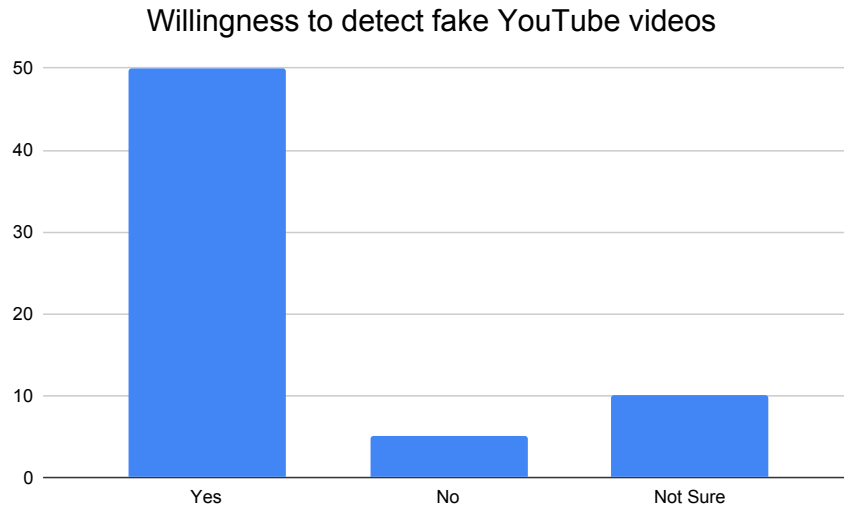


Figure 6.3: Willingness to detect fake YouTube videos

Almost 80% of the users gave their positive opinion on recommending our extension "Bangla Shield" to their family and friends. Among the features most liked or appreciated features were the New information tab and color labeling concept with 49.2% and 67.7% consecutively.

6.3 Thematic Analysis on Feedback

From the survey's point of view, we had to establish some connections for a better understanding of feedback when we performed the thematic analysis.

6.3.1 YouTube Watch Time vs Fake Video

We asked them how they spend their time on YouTube on a daily and weekly basis and how often they encounter fake or misleading content on the platform. Daily YouTube users face more fake videos than everyone, which is 40% and 26 in number among 65 respondents. Figure 6.4 shows their tendency to fall for fake news as YouTube's algorithm keeps suggesting videos according to their watching pattern.

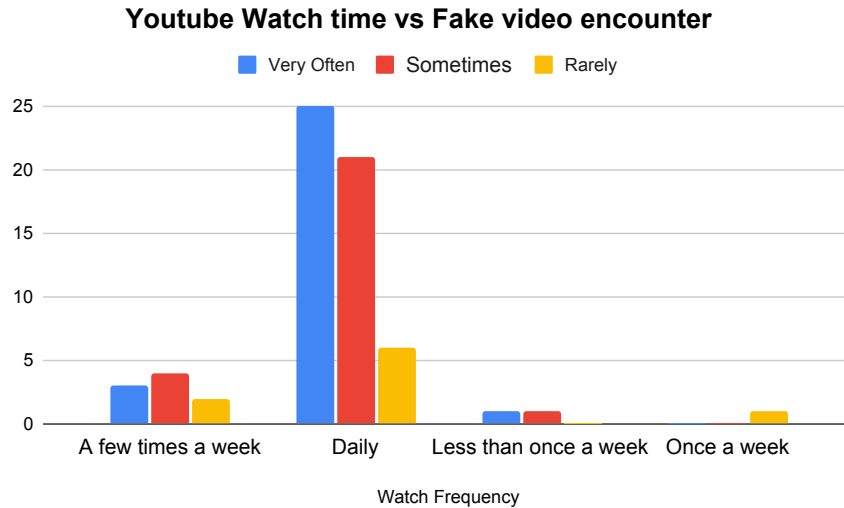


Figure 6.4: YouTube Watch Time vs Fake Video Encounter

6.3.2 YouTube Watch Frequency vs "Bangla Shield" Effectiveness

Among users where watch frequency is divided into 'very often', 'sometimes', and 'rarely' according to the number of users agreeing to use our extension, Bangla Shield. Those who have higher YouTube usage labeled as 'very usually' and 'sometimes' definitely agree to the effectiveness of the extension which, is almost 80% of all respondents. This distribution of responses is shown in Figure 6.5

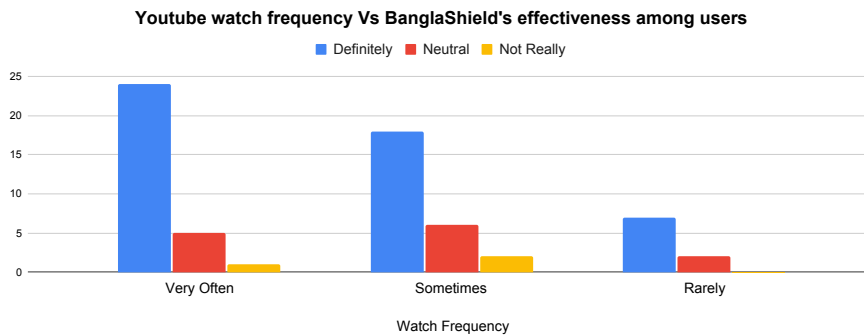


Figure 6.5: YouTube watch Frequency Vs BanglaShield's Effectiveness Among Users

6.3.3 Average Confidence vs Age Group

In this section, we mixed the age groups and the comfortable rate of using extension tools, and respondents showed their confidence level between 1 and 5. When we performed thematic analysis, we found out that the age group 41-50 has the highest confidence in using online extension tools like our creation "Bangla Shield." This age group has a confidence of almost 4 on a scale of 5, whereas 51-60 has an average response or interest and 61-70 has the lowest interest of using extension which is below average, as shown in 6.6.

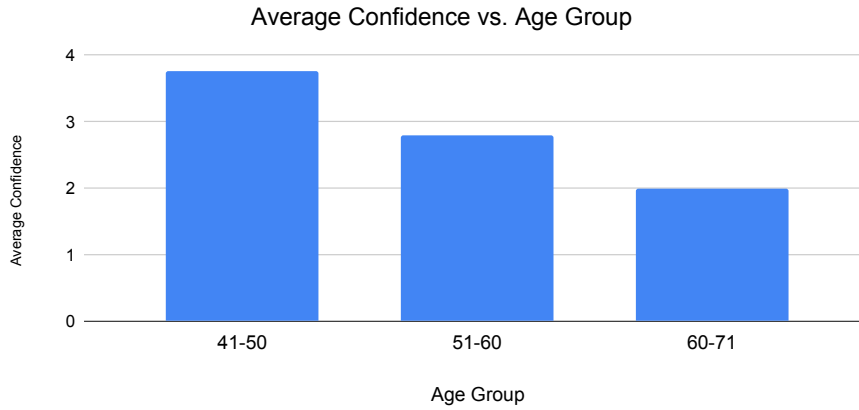


Figure 6.6: Average Confidence vs. Age Group

6.3.4 Extension Clarity vs. Helpfulness of Features

In this section, we mixed up the users who understood our extension work procedure and the helpfulness of the features who having clear and obvious ideas of the extension. Most respondents almost 60% of them found our features helpful and less than 5% of the users felt the features are non-useful, as demonstrated in Figure 6.7

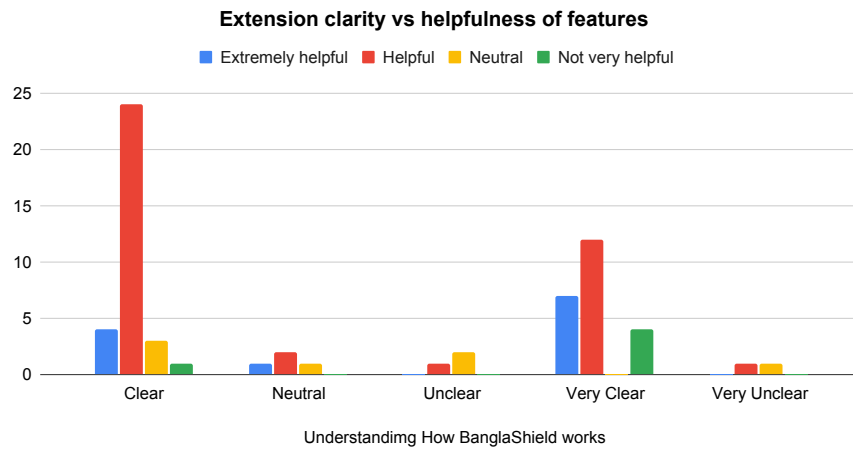


Figure 6.7: Extension clarity vs helpfulness of features

Chapter 7

Discussion

7.1 Challenges Faced by Older Social Media Users

Throughout this research, the findings and results highlight the urgent need to address the challenges and vulnerabilities that older social media users face daily while using YouTube, a platform that has become a primary source of information for many. Over the past years, YouTube has made more information available to everyone than ever before. Every news agency, independent reporters, and everyone switched to YouTube, using this as an information-providing source. Even the entertainment industry has stepped into it, realizing its popularity among users. While this ability to access information from anywhere benefits all to a certain extent, older users comparatively find navigating through YouTube difficult as they face new problems almost every other day. During our extensive research, we've seen these challenges are a result of several factors such as emotional manipulation, lack of digital literacy, or having previous negative experiences on this platform. All of these challenges combine to make our older generation of social media users more and more vulnerable to consuming fake and misleading content, which makes it essential to understand these factors or challenges in detail.

7.1.1 Emotional Manipulation Through Clickbait Titles and Thumbnails

During our research, we have observed that older adults are generally more drawn to any emotional content or something that connects with their sentiment. This content usually draws their attention by using an eye-catching thumbnail and title. These elements are designed in such a way that they can easily provoke curiosity, fear, or hope among audiences. However, such videos have a higher tendency to contain misleading content or irrelevant information. Still, older audiences fail to recognize the pattern of these contents due to their minimal digital literacy, becoming more likely to click on these videos based on the thumbnails and titles alone, which can result in the users being misinformed. This remains a critical factor in spreading fake content among this demographic.

7.1.2 Limited Digital Literacy and Verification Skills

One of the most critical challenges older adults face is their lack of digital literacy; this affects their ability to differentiate between what makes content relative and information and what makes it fake or misinformation. Furthermore, our data has revealed that most older users need to be more informed about how to validate a video's authenticity. They need to be made aware of tools and techniques that may be useful to them. Several methods to identify a forged video include checking video descriptions, source credibility, publication dates, and how users react to that specific content by commenting. This gap in their knowledge makes them more susceptible to falling prey to the irrelevant, fake content circulating on YouTube.

7.1.3 Trust Issues Rooted in Previous Negative Experiences

During our research, users have expressed how their previous negative experiences have impacted their trust in the platform. Participants with prior unpleasant experiences such as scams, misleading videos, and cut or cropped videos without any relevancy have made them more skeptical of the content they view daily. While this level of skepticism can be reasonable, it can create confusion and more difficulty when distinguishing between an actual video with excellent and relevant content and one with no relevancy. This showed us that older generations of social media users in Bangladesh have to overcome several mental obstacles when it comes to detecting actual information from fake.

7.1.4 Lack of Awareness About Content Verification Practices

Based on the results of the quantitative and qualitative analysis, a common issue was the need for older users to be more aware of how to verify content on YouTube. Many participants who participated in our surveys and interviews admitted to not checking the author who published the video when it was posted. This unawareness often results in them consuming older and outdated information fabricated to provide inaccurate information. However, it could be avoided with a better verification process. Moreover, users of this platform were not always familiar with the concept of reporting a video whenever they faced or realized that a video was fake or misinformative, and irrelevant. Several participants did not even have the slightest idea where is the report button or feature in the platform's interface. These challenges occur to our older adults due to a severe lack of awareness when navigating the digital platform safely.

7.2 Implementing Technology for Fake Content Detection

The use of advanced technology has become a necessity to combat the challenges older adults face when browsing YouTube. Using advanced technology offers a hopeful strategy for reducing the spread of misinformation, especially on a platform used by millions of people daily. By using machine learning models and natural

language processing techniques, it is possible to mitigate the challenges older users face. It can be helpful for warning them about potential fake videos or even show them which videos users should totally avoid. In this research, we explore the options and possible solutions of such technological implementations, focusing on the development and performance of various models, their usefulness, effectiveness, and user feedback.

7.2.1 Development of A Browser Extension for Misinformation Detection

To address the issue regarding fake content, we have implemented a browser extension-based system that can fetch and analyze YouTube video comments and classify them. This extension uses NLP models to fetch video comments and cross-check with our database. When users watch a video, the extension processes the comments, and within 10 seconds, it displays a popup with a potential fake percentage of that video. It provides a color-coded visual alert to alert the user about the credibility of that video. This real-time detection gives older adults a decision-making capability to avoid potentially misinformative content.

7.2.2 Performance of NLP and Machine Learning Models (mBERT, Bangla-BERT, Logistic Regression, SVM)

In order to implement our proposed model, ML & NLP models, and training processes were essential to work with YouTube comments, which contain a vast amount of textual data. The system can detect fake comments with reasonable accuracy by utilizing keyword extraction and contextual understanding. As NLP models provide better results for our proposed model, implementing NLP allowed us to filter such misleading and fake comments and provide more accurate results. This helped reduce the chances of older users being exposed to misinformation.

7.2.3 User Feedback on Technological Solutions

Based on user feedback, our BanglaShield extension effectively detects fake YouTube content based on the amount of fake comments. Users have appreciated the pop-up window feature with color labeling made for their clarification. The majority of the participants were willing to recommend it due to its effectiveness in terms of increasing awareness among older adults and also ensuring enhanced safety online. However, participants have expressed different comfort levels when it comes to using technological tools for their daily browsing.

7.3 Effectiveness of a Fake Video Detection Extension

Often, political fake news [19] [57] is very cunning and incorporates true and false information to make it look like it's real. This makes detection harder because normal ways of checking facts might not be able to catch this kind of complicated false information. Some examples can be outdated or irrelevant news, comments made by

politicians taken out of context, cropped videos, influential people’s deepfake videos, etc, which are surfacing more and more recently which affect viewers’ perception of news and sources that can be trusted. In this research work[39], BRENDA is a browser tool to address the problem of false information. BRENDA lets the user first find fact-checking worthy assertions in any news item available online. The user then gets the credibility classification with a complex deep neural network model. The evidence from the model is also shown to the users so they can accomplish all this without ever leaving the news story they are reading on the web page. This research [57]introduces a false news detection approach that analyses self-descriptions from SNS users sharing news URLs and characterizes the bias of words as features. Although most techniques enhance news recognition by integrating content and context variables, model complexity and data collecting still need to be improved. The suggested strategy posits a social network among news-sharing users, relying solely on self-descriptions. To evaluate the proposed strategy, we compared its performance to current methods using different attributes from many datasets. The suggested technique outperformed previous news content feature-based algorithms with an 87.9% classification accuracy on the PolitiFact dataset. Despite being less effective than combining content and contextual information, the suggested technique can be enhanced by combining it with a content-based strategy, as it performs well with contextual features alone [47]. The stated extensions are some examples of the effectiveness of detecting fake videos. Methodologies or approaches may vary, but the main goal is to serve the purpose here and serve as our solution to this vital issue. ” Bangla Shield” is no less as it fetches out the comment section, gives us an idea of fake comment numbers among the top 100 comments using our pre-trained model, and gives a color-labeled popup window detecting fake videos in different levels of fakeness.

7.3.1 Existing Solutions and Bangla-Shield’s Effectiveness Comparison

As mentioned in related works, multiple systems have been proposed to combat fake news. The closest system that’s similar to our proposal is BRENDA [39], which is a browser extension built for Chrome web browser. Users can access BRENDA by clicking on the extension and providing it an article or a selected text, which the extension system can analyze and provide a result mentioning whether the news is True or False with evidence backing its result. By employing a deep neural network model to recognise and evaluate claims on websites, the BRENDA browser addon automates the detection of fake news. In addition to classifying statements as trustworthy or not, this methodology gives users supporting data. BRENDA makes the detection process smooth and effective by enabling end users to quickly verify facts without ever leaving the homepage. Its architecture incorporates tried-and-true methods for handling misinformation at scale, such as deep learning models and hierarchical attention networks. But in our Bangla Shield, we used different NLP models(m-BERT, Bangla-BERT) and ML models(Logistic regression, SVM, XLM-R) by classifying Bangla comments and English-Bangla mixed language over Bangla contents in order to check their credibility and to improve user experience in YouTube. This extension CoReD [49] simultaneously detects a wide range of deepfake videos and GAN images from different generation methods, which is chal-

lenging. Exploring the research works in recent times, very few real-time solutions for fake video detection can be witnessed. Abdulrahman and Baykara (2020) [36] divided their work into three stages- pre-processing, extracting features, and classifiers. The authors used a dataset containing textual data, which they cleaned and preprocessed by removing non-English, removing HTML tags, and applying the ‘stopword’ technique. For extracting features, in they converted the text data into vectors 0 and 1, where we, in our dataset, kept English-Bangla separately and mixed language and performed subjective evaluation observed by different individuals labelling as two ‘Fake’ and ‘Non- fake’. In this study by Waikhom, L.Goswami, R. S. (2019, October) [34], a publicly available LIAR dataset has been used to train a model to detect fake news. However, we made our dataset of over 11000 comments where we prioritized both fake and non-fake labelled comments because to identify the contents’ true self, we need to be precise on the non-fake part, or else only fake labels would bypass any content which itself is fake. Still, its comment section is full of positive comments about the video and misleads any older adults who are consuming the video.

There might be Facebook groups or established websites that can fact-check or detect fake videos. Still, a real-time solution is always better when users can be notified before getting into any loss or falling victim due to a misguided video. BRENDA and this work [48] are two effective examples from our reviewed papers where BRENDA fetches information from the web similar to the video the user is watching and then searches through the web whether across the web it is labeled as false news or legitimate news and another work where we saw fake news detection by social media profile checking with their pre-trained model. Moreover, we, in our research, tried to break the shackles of getting misguided by YouTube content where the target audience is older users. To protect them from getting into any sort of trouble, our extension works in real-time, extracting the comments from the comment section within seven to eight seconds of the video being played, and to prove our statement, we show evidence of 15 fake comments in the new tab after clicking extension icons.

7.3.2 Limitations of Current Detection Models

Future research [19] [57] is encouraged to focus on improving the accuracy of detection algorithms, as well as considering the ethical implications of fake news detection tools to avoid censorship or false positives. For a greater cause or to solve a bigger problem then comes limitations. From the existing literature, what we analyzed in the multilingual barrier where the most efficient solution might fall short due to a language barrier and dataset training according to it. The same goes for BRENDA, CoReD, and our ”Bangla Shield.” So, we targeted our research participants within the Bangla language and were more effective while Bangla content was being consumed. Moreover, real-time solutions are often limited to some extent. For example, we cannot fetch the whole comment section from one YouTube video showcasing all of the fake comments through our pre-trained models. One significant limitation is the lack of digital literacy and technical knowledge gap among older users. Our surveys have shown that even when technological tools are introduced to minimize the spread of misinformation, older adults still struggle due to their lack of familiarity with digital interfaces. Our extension faced several challenges while testing; the accuracy of fake comment detection was not 100%, and there were some false positive

results. Our models are trained on a multilingual NLP model with no significant-sized database. So this can be one of our major drawbacks. Another limitation could be that our proposed model only works on the Google Chrome browser, which is not an optimal solution considering that older Bangladeshi users mostly use handheld devices to browse YouTube.

7.3.3 Potential for Improvement and Future Enhancements

Addressing these limitations requires a diverse method of implementation. Future research should focus on improving the versatility of the system. Training models can achieve this on a more extensive dataset that includes multiple languages and dialects. Moreover, in the future, the need to use an external extension to detect fake content should be extinguished, and platform officials should implement built-in features to make the browsing experience safer for older users of YouTube. Future improvements should also include cross-platform integration to ensure better protection against misinformation. Better real-time detection of fake content across YouTube should be implemented.

Moreover, increasing collaboration between tech companies and policymakers is important to address these issues and develop a permanent solution. Educational campaigns and technological solutions can raise awareness and improve digital literacy. Which can be beneficial to older users to learn the skills and technical abilities one should know to navigate YouTube safely.

Chapter 8

Limitations and Future work

Our results are based on manipulation through fake content on YouTube among older people in Bangladesh. However, our study is also challenged by several limitations. At first, the primary limitation of our study is the target audience and language barriers, as our research is based on older people in Bangladesh. So, we cannot say it represents the entire demographic. It could be more diverse if we took various opinions from people from different countries, but that is challenging because of the language barrier and geography. So, maybe there are some other reasons they might face while using YouTube, which we cannot trace through our data collection. Then, the study could provide even more results and help us better understand.

Another thing was that it was only based on YouTube. There are several other social media platforms, and their systems are different from each other. The methods of spreading false information on those platforms could be different. Also, the techniques of influences and interactions can be more different than YouTube, which we could barely cover in our research.

On the other hand, if we look at our extension, we can see there is a limitation to requesting extension backend server over a large number of videos because our extension only works for 100 videos per day. Moreover, it is a Chrome extension, so there is no option currently available for a mobile-based application and non-chromium web browsers. Also, the current extension version is free for now. But if we introduce a version where we incorporate paid APIs and video fetching is unlimited, then it may be a premium version.

Although these limitations are recognized, several ways of work of further research can be opened upon this work. First, it will be a good initiative to collect information worldwide and do further work for every country. Future studies should also consider different levels of digital literacy and make more sub-groups based on their educational levels and knowledge of using technology. It will help in better understanding on how this influences the vulnerability to fake content.

Another future work could be incorporating other social media platforms to get a larger view of the digital landscape. Each platform has a different and unique way of influencing user behavior. Also, it will be better to come up with a mobile-based approach as the maximum number of users mostly use digital platforms via their

mobile phones.

Finally, future research could be enriched by incorporating advanced machine learning models and interdisciplinary collaboration. The combination of cyber-security, psychology, and sociology experts might help in not only the technical aspects of manipulations but also the intellectual and emotional factors that make older people vulnerable. Other hand, more advanced models with real-time applications can improve the detection and classification of fake content.

Chapter 9

Conclusion

In the modern 21st century, with the growing advancement of technology and the widespread popularity of social media, the risk of fake content getting spread has grown significantly. Especially among the older generation of Bangladesh, this risk is a growing concern, since their lack of digital literacy and awareness makes them more likely to falling for manipulative fake content and potentially spreading it among their family and friends. We have applied various methodologies to investigate this issue, such as online surveys, semi-structured interviews, proxy interviews, and Focus group discussions to identify at what rate the people of the older generation in Bangladesh fall victim to fake content on YouTube. Based on our findings from the survey and interviews, we have observed that older adults lack the skills to use the internet properly; they are not fully aware of the difference between authentic videos with proper sources and fake videos filled with misinformation; also, they have limited knowledge about the dangers of fake news sharing; or they have trouble reasoning clearly. These factors make them more vulnerable to believing or falling for fake content on YouTube. We have also learned that fake videos can influence how older people choose and feel about themselves and others.

Based on our observations, we have proposed a fake and misleading content detection system, which is a browser extension for Chromium-based browsers. To implement this system, we applied Human-Computer Interaction(HCI) methodologies with 5 ML and NLP models to train a custom dataset- among which the mBERT model has proven to be the most accurate with an overall accuracy rate of 92.4%. Our extension ensures a simple design, which makes it easier to use for older social media users. However, there still remain a few limitations, such as the system not being implemented in smartphones and the potential risk of wrongfully identifying a video due to the usage of bot comments. We believe that the system can be perfected with more ideas and development as it gets worked on progressively further in the future, which will help to create a safe and secure digital environment for older users.

Bibliography

- [1] L. E. Yelle, “The learning curve: Historical review and comprehensive survey,” *Decision sciences*, vol. 10, no. 2, pp. 302–328, 1979.
- [2] lipshitz stanley p. and vanderkooy john, “The great debate: Subjective evaluation,” *journal of the audio engineering society*, vol. 29, pp. 482–491, 7/8 Aug. 1981.
- [3] R. B. Cialdini, *Influence: The Psychology of Persuasion*, English. New York: HarperCollins, 1993.
- [4] C. Cortes and V. N. Vapnik, “Support-vector networks,” *Machine Learning*, vol. 20, pp. 273–297, 1995. [Online]. Available: <https://api.semanticscholar.org/CorpusID:52874011>.
- [5] M. F. Kanevski, A. Pozdnukhov, S. Canu, M. Maignan, P. M. Wong, and S. A. R. Shibli, “Support vector machines for classification and mapping of reservoir data,” 2002. [Online]. Available: <https://api.semanticscholar.org/CorpusID:17265429>.
- [6] C. Schmidt, “5.10 the analysis of semi-structured interviews,” *A Companion to*, p. 253, 2004.
- [7] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, “Introduction to the logistic regression model,” 2005. [Online]. Available: <https://api.semanticscholar.org/CorpusID:59226812>.
- [8] J. H. Wu, G. L. Wang, X. M. Li, and S. F. Yin, “Comparison of bp neural network model and logistic regression in the analysis of influencing factors of violence in hospitals,” *Applied Mechanics and Materials*, vol. 50, pp. 964–967, 2011.
- [9] C. Bell, C. Fausset, S. Farmer, J. Nguyen, L. Harley, and W. B. Fain, “Examining social media use among older adults,” in *Proceedings of the 24th ACM conference on hypertext and social media*, 2013, pp. 158–163.
- [10] V. Lopez and D. Whitehead, “Sampling data and data collection in qualitative research,” *Nursing midwifery research: Methods and appraisal for evidence-based practice*, vol. 123, p. 140, 2013.
- [11] A. Talwar and Y. Kumar, “Machine Learning: An artificial intelligence methodology,” en, *International Journal of Engineering and Computer Science*, vol. 2, no. 12, Dec. 2013, issn: 2319-7242. [Online]. Available: <https://ijecs.in/index.php/ijecs/article/view/2261> (visited on 10/15/2024).

- [12] F. Figueiredo, J. M. Almeida, F. Benevenuto, and K. P. Gummadi, “Does content determine information popularity in social media? a case study of youtube videos’ content and their popularity,” in *Proceedings of the SIGCHI conference on human factors in computing systems*, 2014, pp. 979–982.
- [13] D. Khaled, “Natural language processing and its use in education,” en, *International Journal of Advanced Computer Science and Applications*, vol. 5, no. 12, 2014, ISSN: 21565570, 2158107X. DOI: 10.14569/IJACSA.2014.051210. [Online]. Available: <http://thesai.org/Publications/ViewPaper?Volume=5&Issue=12&Code=ijacsa&SerialNo=10>.
- [14] N. Alkış and T. T. Temizel, “The impact of individual differences on influence strategies,” *Personality and Individual Differences*, vol. 87, pp. 147–152, 2015.
- [15] U. Dhavare and U. Kulkarni, “Natural language processing using artificial intelligence,” *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)*, vol. 4, no. 2, pp. 203–205, 2015.
- [16] E. Ferrara, “” manipulation and abuse on social media” by emilio ferrara with ching-man au yeung as coordinator,” *ACM SIGWEB Newsletter*, vol. 2015, no. Spring, pp. 1–9, 2015.
- [17] M. Barthel, A. Mitchell, and J. Holcomb, “Many americans believe fake news is sowing confusion,” *Pew Research Center*, 2016. [Online]. Available: <https://www.pewresearch.org/journalism/2016/12/15/many-americans-believe-fake-news-is-sowing-confusion/%7D>,.
- [18] Z. H. Hoo, J. Candlish, and D. Teare, “What is an roc curve?” *Emergency Medicine Journal*, vol. 34, pp. 357–359, 2017. [Online]. Available: <https://api.semanticscholar.org/CorpusID:16641814>.
- [19] J. Kulshrestha, M. Eslami, J. Messias, *et al.*, “Quantifying search bias: Investigating sources of bias for political searches in social media,” Apr. 2017, pp. 417–432. DOI: 10.1145/2998181.2998321.
- [20] V. Pérez-Rosas, B. Kleinberg, A. Lefevre, and R. Mihalcea, “Automatic detection of fake news,” in *International Conference on Computational Linguistics*, 2017. [Online]. Available: <https://api.semanticscholar.org/CorpusID:27274148>.
- [21] N. Ruchansky, S. Seo, and Y. Liu, “Csi: A hybrid deep model for fake news detection,” in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 2017, pp. 797–806.
- [22] C. Shao, G. L. Ciampaglia, O. Varol, A. Flammini, and F. Menczer, “The spread of fake news by social bots,” *arXiv preprint arXiv:1707.07592*, vol. 96, no. 104, p. 14, 2017.
- [23] A. Singh and R. Shree, “Recognition of natural language processing to manage digital electronic applications,” *International Journal of Advanced Research in Computer Science*, vol. 8, no. 5, 2017.
- [24] H. Ahmed, I. Traore, and S. Saad, “Detecting opinion spams and fake news using text classification,” *Security and Privacy*, vol. 1, no. 1, e9, 2018.
- [25] P. Moravec, R. Minas, and A. R. Dennis, “Fake news on social media: People believe what they want to believe when it makes no sense at all,” *Kelley School of Business research paper*, no. 18-87, 2018.

- [26] A. Quan-Haase and I. Elueze, “Revisiting the privacy paradox: Concerns and protection strategies in the social media experiences of older adults,” in *Proceedings of the 9th international conference on social media and society*, 2018, pp. 150–159.
- [27] T. H. Sandhu and A. R. Itkikar, “Machine learning and natural language processing – a review,” *International Journal of Advanced Research in Computer Science*, vol. 9, pp. 582–584, 2018. [Online]. Available: <https://api.semanticscholar.org/CorpusID:67304307>.
- [28] M. Almaliki, “Misinformation-aware social media: A software engineering perspective,” *IEEE Access*, vol. 7, pp. 182 451–182 458, 2019.
- [29] A. Conneau, K. Khandelwal, N. Goyal, *et al.*, “Unsupervised cross-lingual representation learning at scale,” in *Annual Meeting of the Association for Computational Linguistics*, 2019. [Online]. Available: <https://api.semanticscholar.org/CorpusID:207880568>.
- [30] R. K. Kaliyar, A. Goswami, and P. Narang, “Multiclass fake news detection using ensemble machine learning,” in *2019 IEEE 9th international conference on advanced computing (IACC)*, IEEE, 2019, pp. 103–107.
- [31] C. Parker, S. Scott, and A. Geddes, “Snowball sampling,” *SAGE research methods foundations*, 2019.
- [32] T. Pires, E. Schlinger, and D. Garrette, “How multilingual is multilingual bert?” *ArXiv*, vol. abs/1906.01502, 2019. [Online]. Available: <https://api.semanticscholar.org/CorpusID:174798142>.
- [33] S. Seng, H. Kocabas, M. N. Al-Ameen, and M. Wright, “Poster: Understanding user’s decision to interact with potential phishing posts on facebook using a vignette study,” in *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, 2019, pp. 2617–2619.
- [34] L. Waikhom and R. S. Goswami, “Fake news detection using machine learning,” in *Proceedings of international conference on advancements in computing & management (ICACM)*, 2019.
- [35] H. E. Wynne and Z. Z. Wint, “Content based fake news detection using n-gram models,” in *Proceedings of the 21st international conference on information integration and web-based applications & services*, 2019, pp. 669–673.
- [36] A. Abdulrahman and M. Baykara, “Fake news detection using machine learning and deep learning algorithms,” in *2020 international conference on advanced science and engineering (ICOASE)*, IEEE, 2020, pp. 18–23.
- [37] Y. Agrawal and B. K. Srinivas, “An extensive study for development of browser extensions in secure browser environment,” 2020. [Online]. Available: <https://api.semanticscholar.org/CorpusID:247496794>.
- [38] I. Ahmad, M. Yousaf, S. Yousaf, and M. O. Ahmad, “Fake news detection using machine learning ensemble methods,” *Complexity*, vol. 2020, no. 1, p. 8 885 861, 2020.
- [39] B. Botnevik, E. Sakariassen, and V. Setty, “Brenda: Browser extension for fake news detection,” New York, NY, USA: Association for Computing Machinery, 2020, ISBN: 9781450380164. DOI: 10.1145/3397271.3401396. [Online]. Available: <https://doi.org/10.1145/3397271.3401396>.

- [40] C. Geeng, S. Yee, and F. Roesner, “Fake news on facebook and twitter: Investigating how people (don’t) investigate,” in *Proceedings of the 2020 CHI conference on human factors in computing systems*, 2020, pp. 1–14.
- [41] O. Ngada and B. Haskins, “Fake news detection using content-based features and machine learning,” in *2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, IEEE, 2020, pp. 1–6.
- [42] W. W. Piegorsch, “Confusion matrix,” *Wiley StatsRef: Statistics Reference Online*, 2020. [Online]. Available: <https://api.semanticscholar.org/CorpusID:242242256>.
- [43] J. Ruvika, *Fake news detection*, Accessed: 2024-10-17, 2020. [Online]. Available: <https://www.kaggle.com/jruvika/fake-news-detection>.
- [44] J. Shaikh and R. Patil, “Fake news detection using machine learning,” in *2020 IEEE international symposium on sustainable energy, signal processing and cyber security (iSSSC)*, IEEE, 2020, pp. 1–5.
- [45] U. Sharma, S. Saran, and S. M. Patil, “Fake news detection using machine learning algorithms,” *International Journal of creative research thoughts (IJCRT)*, vol. 8, no. 6, pp. 509–518, 2020.
- [46] A. M. Ud Din Khanday, Q. Rayees Khan, and S. T. Rabani, “Analysing and predicting propaganda on social media using machine learning techniques,” in *2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, 2020, pp. 122–127. DOI: 10.1109/ICACCCN51052.2020.9362838.
- [47] X. Zhou, A. Jain, V. V. Phoha, and R. Zafarani, “Fake news early detection: A theory-driven model,” vol. 1, no. 2, 2020. [Online]. Available: <https://doi.org/10.1145/3377478%7D,%20doi%20=%20%7B10.1145/3377478%7D,>.
- [48] R. Furukawa, D. Ito, Y. Takata, *et al.*, “Fake news detection via biased user profiles in social networking sites,” in *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, 2021, pp. 136–145.
- [49] M. Kim, S. Tariq, and S. S. Woo, “Cored: Generalizing fake media detection with continual representation using distillation,” New York, NY, USA: Association for Computing Machinery, 2021, ISBN: 9781450386517. DOI: 10.1145/3474085.3475535. [Online]. Available: <https://doi.org/10.1145/3474085.3475535>.
- [50] H. Seo, M. Blomberg, D. Altschwager, and H. T. Vu, “Vulnerable populations and misinformation: A mixed-methods approach to underserved older adults’ online information assessment,” *New Media & Society*, vol. 23, no. 7, pp. 2012–2033, 2021.
- [51] D. Trninić, A. Kuprešanin Vukelić, and J. Bokan, “Perception of “fake news” and potentially manipulative content in digital media—a generational approach,” *Societies*, vol. 12, no. 1, p. 3, 2021.
- [52] J. P. Baptista and A. Gradim, “A working definition of fake news,” *Encyclopedia*, vol. 2, no. 1, 2022.

- [53] J. Golzar, S. Noor, and O. Tajik, “Convenience sampling,” *International Journal of Education & Language Studies*, vol. 1, no. 2, pp. 72–77, 2022.
- [54] M. Kowsher, A. A. Sami, N. J. Prottasha, M. S. Arefin, P. K. Dhar, and T. Koshiha, “Bangla-bert: Transformer-based efficient model for transfer learning and language understanding,” *IEEE Access*, vol. 10, pp. 91 855–91 870, 2022. DOI: 10.1109/ACCESS.2022.3197662.
- [55] F. Shahid, S. Kamath, A. Sidotam, V. Jiang, A. Batino, and A. Vashistha, ““ it matches my worldview”: Examining perceptions and attitudes around fake videos,” in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 2022, pp. 1–15.
- [56] F. Shahid, S. Mare, and A. Vashistha, “Examining source effects on perceptions of fake news in rural india,” *Proceedings of the ACM on Human-Computer Interaction*, vol. 6, no. CSCW1, pp. 1–29, 2022.
- [57] T. Simpson and F.-J. Yang, “Some hands-on approaches to fake political news detection,” New York, NY, USA: Association for Computing Machinery, 2022, ISBN: 9781450396912. DOI: 10.1145 / 3556384.3556412. [Online]. Available: <https://doi.org/10.1145/3556384.3556412>.
- [58] J. C. Obi, “A comparative study of several classification metrics and their performances on data,” *World Journal of Advanced Engineering Technology and Sciences*, 2023. [Online]. Available: <https://api.semanticscholar.org/CorpusID:257216116>.
- [59] M. Vaswani, D. Balasubramaniam, and K. Boyd, “A novel approach to improving the digital literacy of older adults,” in *Proceedings of the 45th International Conference on Software Engineering: Software Engineering in Society*, ser. ICSE-SEIS ’23, Melbourne, Australia: IEEE Press, 2023, pp. 169–174, ISBN: 9798350322613. DOI: 10.1109/ICSE-SEIS58686.2023.00023. [Online]. Available: <https://doi.org/10.1109/ICSE-SEIS58686.2023.00023>.
- [60] C. Bisailon, *Fake and real news dataset*, <https://www.kaggle.com/datasets/clmentbisailon/fake-and-real-news-dataset>, Last accessed October 17, 2024, 2024.
- [61] C. Dictionary, *Fake news definition*, <https://dictionary.cambridge.org/dictionary/english/fake-news>, Last accessed October 17, 2024, 2024.
- [62] U. of Hawaii, *Scholarspace at university of hawaii*, <https://scholarspace.manoa.hawaii.edu/items/fd80b2a4-e51d-4704-bd8a-d2377a5d00aa>, Last accessed October 17, 2024, 2024.
- [63] W. L. Hosch, *Youtube*, <https://www.britannica.com/topic/YouTube>, Encyclopædia Britannica. Last updated October 10, 2024, 2024.
- [64] G. M. Insight, *Youtube users statistics 2024*, <https://www.globalmediainsight.com/blog/youtube-users-statistics/>, Last accessed October 17, 2024, 2024.
- [65] S. Issacs, J. Yudelson, and E. Boros, “Classification of fall out boy eras,” *Aresty Rutgers Undergraduate Research Journal*, 2024. [Online]. Available: <https://api.semanticscholar.org/CorpusID:269404308>.
- [66] Kaggle, *Fake news competition*, <https://www.kaggle.com/c/fake-news>, Last accessed October 17, 2024, 2024.

- [67] U. of Oregon Libraries, *Defining fake news*, <https://researchguides.uoregon.edu/fakenews/issues/defining>, Last accessed October 17, 2024, 2024.
- [68] D. Reading, *Why social media sites are the new cyber weapons of choice*, <https://www.darkreading.com/cyberattacks-data-breaches/why-social-media-sites-are-the-new-cyber-weapons-of-choice>, Last accessed October 17, 2024, 2024.
- [69] J. Reshi, “An efficient fake news detection system using contextualized embeddings and recurrent neural network,” *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 8, pp. 38–50, Jan. 2024. DOI: 10.9781/ijimai.2023.02.007.
- [70] G. Support, *Youtube: Uploading and using studio tools*, https://support.google.com/youtube/answer/10834785?hl=en&ref_topic=10833358&sjid=13836841702092060567-AP, Last accessed October 17, 2024, 2024.
- [71] T. O. Oladoyinbo, *Implications of phishing scam activities in adults between age 50-80 in the united states of america*, Unpublished manuscript, n.d.

Chapter 10

Appendix

10.1 Survey Questions

10.1.1 A. General Information

1. Introduction:

- The study focuses on understanding how phishing links and fake content manipulate the elderly population of Bangladesh on Facebook and YouTube. Participants' responses remain confidential and anonymous.

2. Consent:

- Participants are informed that they can withdraw from the study at any time without consequences.

10.1.2 B. Demographics

1. Age Group (Select one):

- 40-50
- 51-60
- 61-70
- More than 70

2. Gender (Select one):

- Male
- Female
- Other

3. Educational Qualification (Select one):

- Primary School Level
- Up to Eighth Standard

- SSC or Equivalent
- HSC or Equivalent
- Undergraduate Degree
- Postgraduate Degree
- Prefer not to mention

4. Location (Select one):

- Capital City
- City Area
- Metropolitan Area
- Rural Area
- Immigrant

10.1.3 C. Social Media Usage

1. Do you use social media apps? (Yes/No)

2. Which platform do you use the most? (Select one):

- Facebook
- YouTube
- Both

3. How do you get your desired information using your devices? (Check all that apply):

- Facebook Posts
- YouTube Videos
- Television and other electronic media
- Google or other search engines
- Online News Portal

10.1.4 D. YouTube-Specific Questions

1. How much time do you spend watching YouTube videos daily?

- Less than 1 hour
- Up to 3 hours
- Up to 5 hours
- More than 5 hours

2. What type of YouTube videos do you prefer to watch? (Check all that apply):

- News and Informative
- Entertainment
- Educational
- Nothing specific

3. Have you ever encountered misleading content on YouTube? (Yes/No)

4. How do you evaluate the credibility of YouTube videos? (Check all that apply):

- The source or author of the video
- The date of the video
- Evidence or references in the video
- Viewer comments or feedback
- Ratings or views of the video

5. Are you aware of the potential for YouTube to be used for spreading false or biased news? (Rating scale of 5)

6. How do you interact with YouTube videos? (Check all that apply):

- I like the videos that I enjoy or appreciate.
- I comment on the videos that I have something to say or ask.
- I share the videos that I think others might like or benefit from.
- I save the videos that I want to watch later or keep for reference.
- I do not interact with YouTube videos in any of these ways.

7. "YouTube content creators spread misinformation via YouTube thumbnails at times." Do you agree with this statement? (Rating scale of 5)

10.1.5 F. Phishing and Security Awareness

- 1. Are you aware of the potential for Facebook or YouTube to be used for phishing (e.g., financial scams or hacking)? (Rating scale of 5)**
- 2. How do you protect your personal information and privacy on YouTube? (Check all that apply):**

- I do not click on suspicious links or ads.
- I use a strong and unique password.
- I do not share personal information in YouTube videos or comments.
- I do not protect my personal information.

10.1.6 G. Perception of Fake Content

- 1. Do you believe news portals on social media verify the truthfulness of their news? (Yes/No/Maybe)**
- 2. How valuable or informative do you find news shared on social media?**

- Very informative
- I would check the source first
- Not believable at all

10.1.7 H. Example-Based Questions

Participants were shown examples of YouTube videos and asked to assess their authenticity or believability. These examples helped gauge the participants' ability to identify fake or misleading content on social media.

1. Example 1: YouTube Video

- Do you think the thumbnail and title of this YouTube video are believable? (Yes/No)
- Do you trust the information presented in the video based on the title and thumbnail? (Yes/No/Not Sure)

2. Example 2: YouTube Video Offering a Free Service of Internet for Lifetime

- Do you think this video claiming "Free YouTube without data" is real or a hoax? (Real/Hoax/Skeptical)

10.1.8 I. Participant Reactions

- 1. How would you react to misleading Facebook posts or YouTube videos? (Check all that apply):**

- Report or flag the video/post

- Comment or warn others
- Ignore or skip the video/post
- Verify the information from other sources
- Do nothing

**2. How do you protect your privacy when using Facebook or YouTube?
(Check all that apply):**

- Avoid clicking suspicious links
- Use strong passwords
- Do not share sensitive information
- Do not protect privacy

10.2 Interview

Question	Information
Select participant age group	Age
Level of education	Education
Residence location	Location
Regular User of YouTube?	Usage
How often would you say you use YouTube?	Usage Pattern
How often do you come across YouTube videos that you believe are misleading or fake?	Frequency of fake video encounter
What type of content do you usually encounter on YouTube?	Type of videos being consumed
Are you familiar with terms such as “fake post” or “fake video”? If yes, kindly elaborate, if no, we can describe it to you.	Understanding the perception
Have you ever encountered content that you later discovered was fake or misleading, and how did you realize it was false?	Encounter with fake/misleading content
Let’s talk about the first video. What do you think about it? <ul style="list-style-type: none"> • Do you think it can be trusted? • How would you describe the title and thumbnail of the video? Was it accurate to the information shown in the video or misleading? • At any point in the video, did you think the video was fake? • What clues made you come up with that thought? Was it something you saw or they said? • How do you judge if a video is fake or not? • Do you look for clues by yourself? Do you ask a family member? Do you look at the source from where it was shared? Or do you just guess if it’s fake or not? 	Understanding the psychological preference
Under video 2’s description several links were stating free money by playing games, etc. Have you ever come across such links? <ul style="list-style-type: none"> • Have you ever heard of phishing links or scam links or hacking links? • Have you ever come across such links? • How did you react to that? • Did anyone share it with you? • Were you beneficial or harmed by clicking on that link? 	Idea of Phishing Link

Question	Information
Potential ways these posts or videos could cause financial loss	Financial Loss
Possible psychological or emotional harm caused by misleading content	Psychological Harm
Experience with falling for fake videos or scams online	Personal Experience
Have you ever reported any fake content or phishing links?	Reporting Fake Content
If yes, what made you take action, and if no, why?	Motivation for Reporting
Are there platforms that are better at preventing fake content than others?	Platforms Comparison
What do you think about the role of tech companies in handling fake content?	Role of Tech Companies
Are there any specific features you believe platforms should introduce to better protect users from fake content or phishing?	Feature Suggestions
Would you be interested in using tools like browser extensions to help identify fake or harmful content?	Interest in Tools
Have you come across any useful tips or practices to avoid falling for scams?	Useful Tips
1. Financial Loss: Examples of financial loss, personal or known experience	1. Financial Loss: Examples of financial loss, personal or known experience
2. Psychological Harm: Nature of harm, emotional impact, personal or known experience	2. Psychological Harm: Nature of harm, emotional impact, personal or known experience
3. Reporting: Platforms, reasons for reporting, suggestions	3. Reporting: Platforms, reasons for reporting, suggestions

10.3 Extension Feedback Survey

1. Are you filling up this form by yourself or on behalf of someone else?

- By myself
- On behalf of someone elderly

2. Select participant age group:

- 41-50
- 51-60
- 61-70

3. How comfortable are you with using online tools like browser extensions?

- Scale 1 (Not Comfortable at all) to 5 (Very Comfortable)

4. How often do you watch YouTube videos?

- Daily
- A few times a week
- Once a week
- Less than once a week
- Rarely

5. In a few words, what kind of content do you usually watch on YouTube?

- Sports
- News & Politics
- Education
- Entertainment

6. If you could easily detect fake YouTube videos, would that be useful to you?

- Yes
- No
- Not sure

7. How often do you come across YouTube videos that you believe are misleading or fake?

- Very often
- Sometimes

- Rarely
 - Never
- 8. Did the screenshots clearly explain how BanglaShield works?**
- Very clearly
 - Clearly
 - Neutral
 - Unclear
 - Very unclear
- 9. How useful is the color-coded labeling system?**
- Scale 1 (Not useful at all) to 5 (Very useful)
- 10. If you see a video labeled as "Orange" or "Red" by BanglaShield, how likely are you to avoid watching it?**
- Very Unlikely
 - Unlikely
 - Neutral
 - Likely
 - Very Likely
- 11. Do you think the detail tab displaying fake comments, view count, and like/dislike information is a helpful feature?**
- Extremely helpful
 - Helpful
 - Neutral
 - Not very helpful
- 12. Do you think BanglaShield will make you more cautious when watching YouTube videos?**
- Yes, definitely
 - Neutral
 - Not really
- 13. Would you recommend BanglaShield to your friends or family? Why or why not?**
- (Opinion based)

14. What did you like the most about BanglaShield, based on the screenshots?

- Pop-up window display
- Color-labeling concept
- Fake 15 comments display
- Fake comment percentage
- Did not like anything