AI SYSTEMS AND CONTENT MODERATION TIKTOK AS A DIGITAL SAFETY PLATFORM IN SHAPING A PLEASANT ENVIRONMENT: A QUALITATIVE APPROACH

Nur Syafiqah Ambran[1], Wan Hartini Wan Zainodin [*2], Muhammad Naim Muhamad Ali[3]
Trust and Safety Department, ByteDance Malaysia, Kuala Lumpur, Malaysia[1]
Faculty of Communication and Media Studies,
Universiti Teknologi MARA, 40450, Shah Alam Selangor, Malaysia[2], MILA University, Malaysia[3]
wanhartini@uitm.edu.my

Abstract

TikTok is an emerging and monopolising social media platform that allows users to create content, watch, share, and discover other users' short videos. Even though TikTok fosters an environment of digital creative expression, harmful content can come in many forms on the platform and is often targeted at minorities. To protect TikTok users, Community Guidelines were created as a benchmark of rules to develop a feeling of safety while having fun on TikTok. Artificial intelligence (AI) technology is used to recognise, filter, and remove videos on TikTok that violate company standards and community guidelines through the moderation process. This study discovers how AI systems and content moderation processes on TikTok can protect community safety from harmful content that goes through the users' For You page (FYP). Through qualitative approach and thematic analysis, 10 informants from the Trust and Safety team were interviewed and all data were transcribed, interpreted, and analysed by NVivo 12 software. Furthermore, the findings show that AI systems and content moderation processes are meticulously crafted to safeguard users' safety by actively monitoring and swiftly removing harmful content, thus fostering a safer online environment for users.

Keywords: *TikTok, Artificial intelligence, Content moderation, Digital safety platform, Social media.*

1.0 Introduction

ByteDance is a Chinese high-tech company that is rapidly entering the overseas market by launching an app named TikTok. TikTok was founded by Zhang Yiming in March 2012 in Beijing, China, and last time it was not recognisable as what we have seen today [19]. Since its inception, TikTok has shown significant ascension as a social media platform and has emerged as the sixth most used social media platform. TikTok allows users to create content, watch, share, and discover other users' short videos on the platform with almost 1 billion active users around the globe using TikTok amongst other social media and almost 14.59 million users aged 18 and above are active TikTok users in Malaysia [18]. After all, TikTok is a place for people to have fun as it fosters an environment of digital creative expression, at the same time, prioritises safety from harmful and disturbing content of its community. Therefore, community guidelines were outlined to protect users from harmful and disturbing content posted by other users on TikTok either intentionally or unintentionally.

TikTok Community Guidelines is a benchmark of rules for creating a feeling of welcome and safe while having fun on TikTok. TikTok has removed almost 81 million videos due to the violation of its Community Guidelines [12]. All content and videos on TikTok undergo compulsory moderation, filtration & and

e-ISSN 2821-3394

removal for the safety of user protection to preserve fundamental rights either by AI system through content moderation processes or by using human labour to moderate content on TikTok [10]. Harmful content comes in many forms and is often targeted at minorities. These include people of different races, ethnicities, religions, LGBTQ+ individuals, and those with disabilities. TikTok must create a safe place for everyone, no matter their background, where they can freely express themselves without fear of harm. Other than that, the removal of content is usually due to heavy or several violations from users, users' reports content, region market policy, and requests from government bodies themselves. The Community Guidelines did not allow any violations as well as nudity or sexual activities, solicitations, use of drugs, dangerous behaviour, minor safety, and more [11]. Contents on TikTok risk unfair censors by the AI systems but little to no research that proves AI systems on TikTok have done a great job as a system.

AI systems and content moderation processes also have often been proposed as the main key tools of technology for identifying and filtering out violation content by TikTok users that curate massive amounts of content [17]. AI systems can recognise images, content, and words in uploaded videos for purposes of categorisation and recommend the content moderation process faster and more effectively [10]. It can process, analyse, and interpret data much faster than humans and at a scale unachievable by human capabilities. The AI systems and content moderation process are based on algorithm systems that can determine users' tailored information distributions by analysing numerous choices and preferences of content accurately based on their search, history, and most watch on TikTok [19]. This system is not only applied on TikTok but several social media platforms have applied where it helps accelerate their daily business process. Due to a shortage of content moderators available at the department, AI systems also help content moderators reduce workload. In the moderation process, this system uses 'algorithms' for the classification of content to support platform moderation [11].

This study attempts to discover how AI systems and content moderation on TikTok can protect community safety from harmful content. These days, most of the content on TikTok might not be safe as every individual has different preferences while creating and watching content. However, AI systems and moderation processes might encounter trouble and leakages which could happen at any time and moment. These issues occur within the convergence of contemporary studies on content moderation, platform regulation, and machine learning fairness, transparency, and accountability [22]. Generally, TikTok's biggest audience is Gen Z who were exposed to technology at an early age thus making them vulnerable to harmful content and needing more preservation [15]. Additionally, this study not only protects adult safety from harmful content on TikTok as well as minor safety.

2.0 Literature Review

2.1 AI systems function on TikTok

Artificial intelligence or AI systems is the replication of human intelligence processes by technology, particularly used by computer systems [12]. Today's AI systems are practically used in the social media field, especially TikTok. AI systems on TikTok allow machines to learn from experience, adapt to new inputs, and execute all human-like tasks [18]. This has resulted in substantial breakthroughs in computer performance in tasks that were previously exclusively attainable for humans [8]. In TikTok, AI systems have a variety of inputs including computer vision and Natural Language Processing (NLP), and detect all content that has been posted by users. This is also reflected in the system which then offers content to the user based on their profile, location, preferences, history, search, and indicated interests. Although this system does not detect a specific interest of the user, the AI system can swiftly identify users' interests.

AI systems on TikTok were set accordingly by ByteDance and the algorithm quickly absorbed individual preferences in recommendation engine data. A recommendation engine filters through data using various algorithms and suggests the most relevant content or videos to users. It first analyses the user's previous activity and then suggests content or videos that the user is likely to watch. On the user's FYP, it will begin with random popular content or videos and become more sophisticated as users spend longer on

e-ISSN 2821-3394

the app and after that follow the recommendation engine. It shows that TikTok's recommendation algorithm repeatedly suggests videos to people that share similar interests or traits and the "targeted catering" technique has significantly increased the number of video views, enabling the quick diffusion of high-quality content [13]. Meanwhile, the recommendation engine would categorise the content with subjective tags and conduct natural language processing and synthesis [23]. The categories of subjective tags include country setting, age, language preferences, and device type after allocation, the recommendation engine would rate content based on the potential of a user's interest.

AI systems can be tested for their transparency, safety, accountability, comprehensibility, and ethical ideals on social media platforms like TikTok [21]. Transparency and ethical ideals can be seen when several contents related to sexual issues are not monitored tight by the AI systems and it is also suggested to individuals with similar interests to that. That means all those types of harmful content still existed on the platform. On TikTok, some harmful content that does not hit serious policy violations is still published. Users can choose to continue to watch the content or skip it anyway. In order to protect their most vulnerable TikTok users, they have strengthened the security procedures to avoid any disruption in the future [21]. Other than that, TikTok still improves its system from all harmful content, data, and security. In a study show that the world's most popular social media platforms including Instagram, Pinterest, and TikTok, have increased and strengthened their safety efforts by collaborating with independent experts [23]. Hence, this study will observe more on how AI systems can prevent all harmful content and other disruptions from going through users' FYP.

2.2 Content moderation removal processes

All major social media services participate in moderation processes that fall along the spectrum, then discovered a variety of cases through his research on interviews, documentation, and by his service testing of the content moderation process [16]. Content moderation removal processes sometimes can lead to users' emotions of marginalisation, exploitation, distrust, and unwelcomeness on the platform [10]. During the process of removing content, AI systems in content moderation can cause extra issues due to a lack of process for specific user content, video, and their current or past activity [17]. However, because TikTok used automated processes of AI systems and is still in its early stages, the platform may struggle to appropriately identify content deserving of removal. According to Gaffney; a Security Consultant for F-Secure in a Trusted Reviews interview, stated that the content moderation process will inevitably, there surely will be teething problems that can cause some harmful content to slip through and also false positives to occur, it will take time for TikTok's algorithms to become refined enough to prevent the misclassification of video content [27]. Several studies show that there are problems with the moderation process, this study will look into how content moderation removal works.

2.3 Transparency in content moderation processes

Transparency in the moderation process matters, and lack of transparency in moderation processes can reduce the comprehensiveness of content regulation, diminishing users' trust in social media networks. With lacking platform transparency on moderation processes, TikTok users' content or video might be reduced to basic coding faults or platform exclusions of cultural communication nuances. According to [10], transparency in content moderation has crucial emotional consequences for TikTok content creators in terms of the content they make and their perspective toward the platform. However, he believes that platform transparency should not be prioritised over the knowledge and insight that platform users, particularly those from underprivileged backgrounds. Meanwhile, according to [28] a Chief Operating Officer at TikTok, they strengthened transparency by launching Transparency and Accountability Centers two years ago where they allowed experts to access moderation practices and information and they also started publishing transparency reports; expanded in each Community Guidelines Enforcement Report. In

e-ISSN 2821-3394

general, this approach is to increase transparency and explain to authorised users why their content or video was deleted.

2.4 User safety and well-being on TikTok

One of the most important issues of this study is users' safety and well-being while scrolling and enjoying TikTok. The safety and well-being of TikTok users are considered their top priority on the list. TikTok has cooperated in order to maintain users' safety and support amidst mental health challenges from harmful content [7]. Social media and well-being highlighted those various psychological processes, such as upward social comparison or fear of losing out on viral content, are related to psychological emotion and may have a significant impact on TikTok users' life in general [23]. Meanwhile another study stated that TikTok could be a powerful medium for educating young people about health-related information and official content given by the government or other media agencies, but it can also be a source of harmful health content, such as e-cigarette smoking that can harm users' [29].

Both studies show users' safety and well-being from positive to negative impacts which are related to psychological issues after a day spend on TikTok. According to [33], most parents were concerned about their children's exposure to harmful content and social media activity, and being approached or exploited by random people on the internet frequently consider strict privacy restrictions as critical to preventing damage. Based on the previous study, we can see that the users' safety and well-being did not guarantee in the future. Therefore, this study will dig more into how TikTok continues to keep its users safe.

3.0 Methodology

3.1 Informants

This study uses qualitative methods with a meticulous and subjective approach to investigate AI systems and content moderation and its role in protecting TikTok users. The aim is to imbue life experiences with significant value. A phenomenology approach employed in this study can uncover explicit values that reflect axiological assumptions where the informants were allowed to express their thoughts, feelings, and experiences in their own words, free of the constraints imposed by fixed-response questions [38]. This study was carried out in the Klang Valley area where ByteDance Malaysia is located. This study used purposive sampling and 10 moderators were picked based on criteria including from the Trust and Safety team with experience of more than one year in the moderation field. [9] recommends a sample size ranging between 5 and 25 participants, emphasising that this range is generally adequate for achieving the goals of phenomenological inquiry. Meanwhile, according to [32], a qualitative sample size of 10 people is sufficient for sampling among a homogeneous population. Meanwhile, Furthermore, data saturation was achieved during the 10th interview when the researcher noticed that no new theme emerged. It is because to develops a strong relationship with participants, resulting in more natural talks and better data [25].

Table 1: Informants' Background

Informant	Years of experiences as content moderator	Background of studies
CM1	2 Years	Business administration
CM2	3 Years	TESL
CM3	3 Years	Language studies
CM4	3 Years	Forensic
CM5	2.5 Years	Business administration

e-ISSN 2821-3394

CM6	2.5 Years	Business administration
CM7	3 Years	Human Resources
CM8	2.5 Years	Accounting
CM9	3 Years	Banking
CM10	3 Years	Physiotherapy

3.2 Data Collection

The interview protocol has been developed as it is an in-depth interview. In designing the interview protocol, the researcher creates open-ended questions. Open-ended questions give informants more time and space to open up and share more details about their experiences [15]. An interview protocol is a valuable tool that provides the researcher with the necessary consistency [33]. After the interviews, all the data will be transcribed, interpreted, and analysed using NVivo 12 which assists in organising, analysing, and visualising mixed media and unstructured data. By providing tools for classifying, sorting, and arranging data in ways that allow the researcher to identify themes, patterns, and cloud charts of the study [25].

4.0 Result and Discussion

This part will discuss the result of how TikTok content moderation as a digital safety platform can create a safe environment for society. The theme has been created and divided into eight; behaviour tracker, block inappropriate content, classified label, community guidelines, equitable system, human team support, reported content, and safety features will be discussed further in this part.

Figure 1: Word cloud of themes of AI systems and content moderation TikTok as a digital platform for shaping a pleasant environment.

Based on the above figure, this study attempted to prepare a word cloud using excerpts. The word cloud is used for a better understanding of this study data. Moreover, word clouds visually display common keywords found in qualitative data. The larger the word, the more often it appears in the data. They give a quick glimpse into the main topics or trends present in the information.

4.1 Behaviours Tracker

e-ISSN 2821-3394

AI systems are the type of technology that can learn and adapt to data input and mimics human intellect by tracking users' daily activities. The system can understand better what each user wanted to watch by following their browsing or history and it utilises a range of technology [22]. As mentioned by CM1, *'This system track based on your browsing, history, activity, and habits too and it will track user behaviour and this is what it means to track every behaviour of user activity'.* Meanwhile, CM2 stated that *'The algorithms quickly learn and track individual preferences, as they capture not only the users' "likes" and comments, but how long they watch each preference video in a day'.* Based on the statement, AI systems execute predictive analytics on the users' indicated interests and preferences, as well as user tracking and behaviour based on user profiling [14].

Furthermore, the recommendation videos system is relatable to the behaviours tracker and is highlighted on the top of users' FYP flagship feature. CM1 mentioned, *'This technology optimises its content creation, curation, and recommendation. As you can see how it works it's magical, especially in tracking every behavior of what users watch on TikTok'.* Based on the statement above, he believes that content creation and curating of users' daily activity will lead to videos recommended on TikTok. Another study shows that the recommended system is based on video rank on a variety of characteristics depending on the user's behavior on the app, including optimising for items you've indicated may not be of interest [22]. A stream of videos on TikTok is personalised to users' interests, which makes it easier to find the content you like. Driven by a recommendation algorithm, it offers content to each user that is likely to be relevant to them. It also crossovers with popular videos, each FYP stream is distinct and tailored to specific users.

4.2 Block Inappropriate Content

Block inappropriate content is an important feature that was set in data AI content moderation to ensure that the content is safe for users to watch. This program is very valuable since it can filter inappropriate content before TikTok users see it. The core focus of inappropriate content not only monitors sexually explicit material but various issues such as violence, both physical abuse and political related to extremism or terrorism, hostile, insulting, incorrect or inaccurate information, and more, which have lately been included in the database [17]. CM3 stated that *'TikTok's main priority is to keep the platform safe from any suspicious and disturbing content and it is also important for community safety. AI helps by combing thru the platform by capturing and blocking immediately any suspicious content and taking down the video'.* This shows that TikTok AI systems prioritise users' safety by removing harmful content from being on the platform before can be seen by other users. It needs to be removed immediately as it is might affect certain users' psychological issues. Most of the videos were removed because of the violation of Community Guidelines as it is a benchmark of this platform.

Moreover, the removal of content on TikTok is because violation of Community Guidelines and will be automatically removed by AI systems if it identifies and flags a potential violation of policy. This moderating process is defined as being focused on strict organisational goals. Sometimes, the app may restrict the accessibility of your content for reasons such as spamming your account or uploading harmful content. As mentioned by CM7 *'It quickly recognises human needs to filter, remove, block, and detect inappropriate content that goes on TikTok and also the one that does not meet TikTok's Community Guidelines. This inappropriate content will be blocked automatically by the AI system'.* According to [5], this computer-based mechanism could only identify a video with potentially dangerous content, while a squad of human moderators was in charge of reviewing and, if required, removing the video. Without being said, the app only eliminates accounts that breach the Community Guidelines, which describe what you can and cannot do and upload on TikTok.

e-ISSN 2821-3394

4.3 Classified Label

Generally, TikTok has used a machine learning algorithm to provide rapid and informative data on their platform. In Community Guidelines, there are several violation categories themes and sub-themes that are depicted on the TikTok website, and the violation content will be classified by those categories. Apart from that this classification is to systematise the violation content from the system and make human moderators' work easier [28]. CM9 stated, *'the first filtration on the thousands of contents posted every single day is to prevent leakage that could impact society and minors' safety'.* After AI content moderation filters content, it will go to the moderator for the second filtration to filter more detail and classify which label the content should be. From the quote, this categorisation of violation will go to each theme that is related to the user's content by the automatic system as stated in Community Guidelines. Additionally, all these general categorisations will also go to the human moderator to evaluate in detail categorisation before it goes to the public or is removed.

TikTok clarifies that automated will be restricted for content categories where this technology is most accurate and that it will begin with breaches involving minors' safety, adult nudity, sexual activities, violent and graphic material, illicit activities, and regulated goods [12]. Besides that, CM10 stated, '*these AI systems will filter and separate all inappropriate content into various categories label such as mutilation, blood, death and so on in assisting human moderators by filtering suspicious content for human review'.* Thus, it prevents content moderation teams from having to go through all the content reported by users and reduces human exposure to disturbing content. In order to remove this content and organise it into various categories, it needs to identify overall content, sound, live stream, pictures, comments, hyperlinks, or captions that violate Community Guidelines. Furthermore, AI systems will temporally or permanently remove all high-risk violation content that might not be safe to watch the public.

4.4 Community Guidelines

Community guidelines are a set of policies issued by each social media platform to ensure a level of behavior anticipated on the platform in order to foster a safe environment for users to communicate and have fun [4]. They are intended to assist users in understanding what is required of them in the community space, as well as significant dos and don'ts on TikTok. All the violation content that hits Community Guidelines on TikTok will be automatically removed from the platform. CM9 mentioned that *'there is certain action that prohibits such as something that brings danger or harm including minor safety. This is also not limited to humans, we even control animated or digital content on the platform'.* Community guidelines are a place to ensure user safety while scrolling TikTok but if someone is repeatedly against the guidelines, their account could be banned from TikTok'. Based on the statement, TikTok promotes safety, uniqueness, inclusiveness, and sincerity. They also encourage users to express their uniqueness and audiences to participate by what drives them to feel a safe atmosphere that allows each user to do content freely. Meanwhile, the head of Trust and Safety TikTok, Cormac Keenan (2022) said that they utilise a combination of AI technology and humans to detect and eliminate violations of their Community Guidelines, and they will continue to train automated systems and safety teams to enforce regulations [1].

4.5 Equitable System

AI content moderation can process data considerably quicker than people, allowing it to detect patterns much faster, and it can also review significantly bigger datasets than humans, allowing it to reveal patterns that humans would just miss [34]. It can aid in the development of prediction models and algorithms for data processing and understanding the potential outcomes of various trends and events that happened on TikTok. The reason this system existed is to reduce human error and increase accuracy and precision. As mentioned by CM7, *'The videos mostly follow our Community Guidelines, and AI systems have done a great job removing inappropriate content. People nowadays understand the process of reporting content if they find it uncomfortable to be viewed by society and also helps our moderation*

e-ISSN 2821-3394

process. As a result TikTok specifically has no issues with the system'. AI systems also reduce the time required to accomplish a task, enable multitasking, and reduce the demand for current resources. Aside from that, this system enhances user interaction and content creation in a safe environment.

On top of that, the AI system is emotionless and extremely realistic and reasonable in its approach, and free from bias, resulting in a more accurate decision-making algorithm. It is also has been revealed that AI systems have attained ideal accuracy and speed in identifying harmful content on social media platforms [12]. Moreover, CM8 stated that *'The responsibilities of AI systems is a big, as the moderator is considered as the second after the process. That means the system already took half the work of the moderator. The system just works fine with no issues with detection, it helps with the daily task as a content moderator. So far we have no big issue to deal with, even though there is an issue TikTok still can fix it'.* Furthermore, the possibility of inaccurate detection content of AI systems is less. With a great system, AI content moderation not only protects users' safety but includes data privacy, information, brand image, the reputation of the business, and more from cyber criminals' attacks on TikTok [30].

4.6 Human Team Support

Since AI content moderation is an equitable system, it would have no issues involving the system. Although AI works better, moves quicker, and lifts heavier weights than humans because it is more intelligent, it can accomplish practically anything. Based on the data gathered, moderators can be considered as a support to AI content moderation on TikTok as stated by CM6, *'The moderators come into action as a support and process of AI content moderation to tag, remove, and filter those not flagged by the system. Besides that, there aren't pressing issues with detection'.* Moderators can support AI content moderation in their viewing process and it helps businesses to expand more quickly given their resources. The capacity to effectively detect and rapidly delete improper content is critical for community safety. Overall, it just makes the process quick. However, technology cannot progress without the assistance of people. Engineers and scientists are needed to design and test AI systems for the technology to advance. As a result of the informant's quote, humans and AI systems are not interchangeable, and AI systems and content moderation processes cannot exist without humans.

To maintain its platform free from harmful content, TikTok employs both AI systems and human moderators to support the safety team to smooth the content moderation processes. Human moderators are the first response and management who make TikTok a safe environment for all of us. It is essential in a business where TikTok hires human moderators to support the AI system in filtering content as the labor cost of a moderator is cheaper than an AI system. CM10 mentioned that *'If leakage happens from AI content moderation, we human moderators might control it or any management related will protect the violation from going out to the platform. It is also will go to moderators and the moderator will filter the content again to make sure no leakages happen'.* Based on the statement, CM10 believe that human moderator helps AI system in filtering content to avoid leakages. When TikTok content moderators flag a piece of content for removal, they are not just removing it. They are also gathering information about the exact regulations that it breaches, which will be used to improve the platform's machine-learning systems to better in the future [23].

4.7 Reported Content

Reported content is content that has violated Community Guidelines. All those reported content will be reviewed by the Trust and Safety team to see whether the content violated any of the Community Guidelines. There are various causes of reported content including nudity, bullying, terrorism, hacking, hate speech, impersonation, bullying, illegal content, and any things that violate TikTok's standards could be reported. If the system does not remove any content that feels disturbing to be seen by the public and is not even safe for society to review, users can report the content as mentioned by CM7, *'We always encourage users to report content that they believe violates TikTok's guidelines because we want the best*

e-ISSN 2821-3394

for users where they can scroll the app without encountering any inappropriate. This is the best solution that we always give to users. He continues with the process of reporting and states 'Users can read how to report violated content on our website or go to the TikTok help centre to read how to report the video. The bottom of the TikTok app has a button that they can report. They just need to follow instructions and they are done reporting videos. From the quote, the informant keeps emphasising that TikTok always encourages users to report the content because they want the greatest experience for all TikTok users without facing any consequences.

Moreover, CM8 stated that *'Human moderators can detect the leakage and can retrieve the content from the platform itself. Users also will report immediately to us if they see any disturbing content. So that is how we detect from the platform. Users always help us by doing this'.* Not to say about other things but the user is very good in terms of reporting videos that violate policy. Reported content between five and six reports can result in a permanent account suspension and this is a significant amount when compared to other sites, such as Facebook, Instagram, and YouTube [11]. The suspension depends on how seriousness and frequency of their violation. Temporary bans typically take about 24-48 hours of the ban and a permanent ban will be permanently banned the account from TikTok itself. According to [1], TikTok believes that individuals will continue to discover methods around the algorithm because 60% of those who received only a warning have done the same crimes again. From this TikTok will make sure Community Guidelines and policies extra tighten in the future if users continue to violate policies.

4.8 Safety Feature

AI system and content moderation process can be considered a great system that has ever been developed by engineers. Online safety and data privacy issues are always top of the list. TikTok has provided several data privacy and information control options that you can configure from your phone to ensure a more secure experience on TikTok. CM2 did mention in the interview session *'TikTok takes the safety of users as a priority. Besides that, it constantly working towards improving mechanisms by introducing industry-leading safety features such as increased privacy settings, in-app reporting, comments filter, and strict Community Guidelines'.* From the quote, TikTok always prioritises the safety of the community on the platform such as users' data privacy and information. Besides, TikTok has a range of privacy and safety options that allow users to control who may contact and comment on their postings and profile. This also includes TikTok's privacy policy which this the app takes anything possible to safeguard its users' stored personal data, including utilising encryption.

During the interview session, CM5 mentioned that minor safety is also important as they a still small to understand what dos and don'ts of TikTok. Several violations under minor safety are Flattery, demands for interaction on or off the platform, requests for personal information, solicitation of minor sexual assault content, sexual solicitations or statements, and gift-giving are all the elements. He stated that *'AI systems are important for society to make sure the platform is safe for everyone as a whole and to maintain minor safety and also user mental health. Minor safety is guidelines that create for minors. This specific guideline is just for minors and if the content is someone under age we will remove it under this. So yes, we also prioritise minors in TikTok we always monitor this type of content and most of our consumer is from youngsters'.* In order to protect minor safety, TikTok has a specialised design parental supervision tool in order to protect minor safety on the platform while enjoying watching videos on TikTok. According to [33], TikTok released a package of digital well-being and safety tools that provide parents with more monitoring and restriction alternatives unprecedented move by a social media site, the consequences of which are unknown. A safety feature for minors or teenagers under the parental supervision tool is called a 'Restricted Mode' or screen limits time on apps. In Restricted Mode, a parent must include access to their teen's phone in order to input a code that allows only the parent to switch the app out of Restricted Mode or enable use over the allocated screen time. With this type of safety feature on TikTok, users, and parents can assure the safety of TikTok.

e-ISSN 2821-3394

5.0 Conclusion

In general, this study seeks to provide a fundamental of AI systems and the content moderation process on TikTok as a digital safety platform in shaping a pleasant environment for all users. The result has shown a positive outcome. TikTok is in charge of formulating and implementing rules and safety practices aimed at keeping people safe across all countries in the world. The safety of users is a top priority at TikTok. Content submitted to TikTok is first processed by AI systems, which detect and flag potential policy breaches for further assessment by a member of the safety team. If there is a violation, AI systems will instantly erase the video, but they still have the option to appeal the video. If no violations are found, the video will be posted and seen by others on TikTok. Hence, TikTok is continuously striving to enhance our systems by presenting industry-leading safety features like enhanced privacy policies, in-app reporting, comments filtering, strict Community Guidelines, a safety resource centre, and a favourable in-app atmosphere for our users to display their creativity.

However, several studies emphasise the uses of method computer vision, Natural Language Processing (NLP), and metadata in classifying the content on TikTok but it is not mentioned by informants. According [3], they suggested technique of moderation processes begins with evaluating the content using three factors computer vision, NLP, and metadata. This computer vision is a deep learning process that uses neural networks to decipher images within a photo or video. The algorithm is backed by a dataset of millions of labelled images that allows the algorithm to recognise new images based on specific traits and characteristics. It enables the algorithm to see and understand the content of the videos being created. Besides, this NLP is used to translate and describe the audio content and after that, the final stage is metadata; detection of the content caption, hashtags, etc.

6.0 References

[1] Ahmed, A. (2021). TikTok enables AI-based automated removals to prevent harmful content from being exposed at all; Has a false positive rate of 5 percent, Digital Information World, https://www.digitalinformationworld.com/2021/07/tiktok-enables-ai-based-%20automated.html

[2] Al-Ghamdi, L. M. (2021). Towards adopting AI techniques for monitoring social media activities. Sustainable Engineering and Innovation, 3(1), 15-22.

[3] Ali, M. Y., Naeem, S. B., & Bhatti, R. (2020). Artificial intelligence tools and perspectives of university librarians: An overview. Business Information Review, 37(3), 116-124.

[4] Anderson, K. E. (2020). Getting acquainted with social networks and apps: it is time to talk about TikTok. Library hi tech news.

[5] Cobbe, J. (2021). Algorithmic censorship by social platforms: Power and resistance. Philosophy & Technology, 34(4), 739-766.

[6] Creswell, J. W. (2009). Research Design Qualitative, Quantitative, and Mixed Methods Approaches(3rd ed.). Thousand Oaks, CA Sage Publications.

[7] Díaz, Á., & Hecht-Felella, L. (2021). Double standards in social media content moderation. Brennan Center for Justice at New York University School of Law. https://www.brennancenter.org/our-work/research-reports/double-standards-socialmedia-content-moderation

[8] Drootin, A. (2021). " Community Guidelines": The Legal Implications of Workplace Conditions for Internet Content Moderators. Fordham L. Rev., 90, 1197.

[9] Du-Harpur, X., Watt, F. M., Luscombe, N. M., & Lynch, M. D. (2020). What is AI? Applications of artificial intelligence to dermatology. British Journal of Dermatology, 183(3), 423-430.

[10] Fernandes, M. B., (2022), Making Sense of Digital Content Moderation from the Margins, e Virginia Polytechnic Institute, and State University.

e-ISSN 2821-3394

[11] Gerrard, Y. (2022). Social Media Moderation: The Best-Kept Secret in Tech. In The Social Media Debate (pp. 77-95). Routledge.
[12] Gongane, V.U., Munot, M.V. & Anuse, A.D. (2022), Detection and moderation of detrimental content on social media platforms: current status and future directions. Soc. Netw. Anal. Min. 12, 129
[13] Hassani, H., Silva, E. S., Unger, S., TajMazinani, M., & Mac Feely, S. (2020). Artificial intelligence (AI) or intelligence augmentation (IA): what is the future?. Ai, 1(2), 8.
[14] Heldt, A., & Dreyer, S. (2021). Competent third parties and content moderation on platforms: Potentials of independent decision-making bodies from a governance structure perspective. Journal of Information Policy, 11, 266-300.
[15] Hoffmann, E. A. (2007). Open-ended interviews, power, and emotional labor. Journal of contemporary ethnography, 36(3), 318-346.
[16] Kaplan, A. (2020). Artificial intelligence, social media, and fake news: Is this the end of democracy?. IN MEDIA & SOCIETY, 149.
[17] Keenan, C. (2022), Strengthening our policies to promote safety, security, and well-being on TikTok, TikTok, https://newsroom.tiktok.com/en-us/strengthening-our-policies-to-promote-safety-security-and-wellbeing-on-tiktok
[18] Kemp S. (2022). Digital 2022: Malaysia — DataReportal – Global Digital Insights. Retrieved May 14, 2022, from https://datareportal.com/reports/digital-2022-malaysia
[19] Llansó, E. (2020). One in a Series of Working Papers from the Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression Artificial Intelligence, Content Moderation, and Freedom of Expression. www.annenbergpublicpolicycenter.org/twg
[20] Liu, H. C. (2020). Artificial intelligence stomatology. Zhonghua kou Qiang yi xue za zhi= Zhonghua Kouqiang Yixue Zazhi= Chinese Journal of Stomatology, 55(12), 915-919.
[21] Ma, Y., & Hu, Y. (2021). Business Model Innovation and Experimentation in Transforming Economies: ByteDance and TikTok. Management and Organization Review, 17(2), 382–388. https://doi.org/10.1017/MOR.2020.69
[22] Matamoros-Fernandez, A., Gray, J. E., Bartolo, L., Burgess, J., & Suzor, N. (2021). What's "up next"? Investigating algorithmic recommendations on YouTube across issues and over time. Media and Communication, 9(4), 234-249.
[23] Medvedeva, M., Vols, M., & Wieling, M. (2020). Using machine learning to predict decisions of the European Court of Human Rights. Artificial Intelligence and Law, 28(2), 237-266.
[24] Mishre, R. P., & Picek, S. (2021). How can regulation enhance transparency of AI facilitated content moderation.
[25] Mortelmans, D. (2019). Analyzing qualitative data using NVivo. In The Palgrave handbook of methods for media policy research (pp. 435-450). Palgrave Macmillan, Cham.
[26] Montag, C., Yang, H., & Elhai, J. D. (2021). On the psychology of TikTok use: A first glimpse from empirical findings. Frontiers in public health, 9, 641673.
[27] Novick, G. (2008). Is there a bias against telephone interviews in qualitative research?. Research in nursing & health, 31(4), 391-398.
[28] Pappas, V. (2022). Strengthening our commitment to transparency. TikTok Newsroom. TikTok. https://newsroom.tiktok.com/en-gb/strengthening-tiktoks-commitment-to-transparency
[29] Papcunová, J., Martončik, M., Fedáková, D., Kentoš, M., Bozogáňová, M., Srba, I., ... & Adamkovič, M. (2021). Hate speech operationalization: a preliminary examination of hate speech indicators and their structure. Complex & Intelligent Systems, 1-16.
[30] Parimala, A., Vijayalata, Y., & Deepa, R. A. (2022, April). Survey on Image Authentication and Privacy in Public Networks. In 2022 6th International Conference on Trends in Electronics and Informatics (ICOEI) (pp. 01-09). IEEE.

e-ISSN 2821-3394

[31] Ryles, G. (2021). TikTok now uses AI for user violations – here's what it means for you. Trusted Reviews. https://www.trustedreviews.com/news/tiktok-now-uses-ai-for-user-violations-heres-what-it-means-for-you-4154531

[32] Sandelowski, M. (1995). Sample size in qualitative research. Research in nursing & health, 18(2), 179-183.

[33] Savic, M. (2021). Research Perspectives on TikTok & Its Legacy Apps| From Musical. ly to TikTok: Social Construction of 2020's Most Downloaded Short-Video App. International Journal of Communication, 15, 22.

[34] Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. Nature medicine, 25(1), 44-56.

[35] Wawrzuta, D., Jaworski, M., Gotlib, J., & Panczyk, M. (2021). Characteristics of antivaccine messages on social media: systematic review. Journal of Medical Internet Research, 23(6), e24564.

[36] Weimann, G., & Masri, N. (2020). Research note: Spreading hate on TikTok. Studies in conflict & terrorism, 1-14.

[37] Yadalam, T. V., Gowda, V. M., Kumar, V. S., Girish, D., & Namratha, M. (2020, June). Career recommendation systems using content based filtering. In 2020 5th International Conference on Communication and Electronics Systems (ICCES) (pp. 660-665). IEEE.

[38] Zainodin, W. H. W., Ibnu, I. N., Ambikapathy, M., & Bakar, Z. A. (2022). Democratically speaking: YouTube as a voice of freedom among Malaysian Gen Y. SEARCH Journal of Media and Communication Research (SEARCH), 167.

[39] Zhang, Z., & Gupta, B. B. (2018). Social media security and trustworthiness: overview and new direction. Future Generation Computer Systems, 86, 914-925.Zhang, M., & Liu, Y. (2021). A commentary of TikTok recommendation algorithms in MIT Technology Review 2021. Fundamental Research, 1(6), 846-847.

e-ISSN 2821-3394
