Journal of information and communication convergence engineering
http://jicce.org

J. Inf. Commun. Converg. Eng. 21(1): 75-81, Mar. 2023 Regular paper

Implementation and Evaluation of Harmful-Media Filtering Techniques using Multimodal-Information Extraction

Yeon-Ji Lee, Ye-Sol Oh, Na-Eun Park, and Il-Gu Lee*

Department of Future Convergence Technology Engineering, Sungshin Women's University, Seongbuk 02844, Korea

Abstract

Video platforms, including YouTube, have a structure in which the number of video views is directly related to the publisher's profits. Therefore, video publishers induce viewers by using provocative titles and thumbnails to garner more views. The conventional technique used to limit such harmful videos has low detection accuracy and relies on follow-up measures based on user reports. To address these problems, this study proposes a technique to improve the accuracy of filtering harmful media using thumbnails, titles, and audio data from videos. This study analyzed these three pieces of multimodal information; if the number of harmful determinations was greater than the set threshold, the video was deemed to be harmful, and its upload was restricted. The experimental results showed that the proposed multimodal information extraction technique used for harmful-video filtering achieved a 9% better performance than YouTube's Restricted Mode with regard to detection accuracy and a 41% better performance than the YouTube automation system.

Index Terms: Harmful media, Media filtering, Multimodal information, OCR, Video filtering

I. INTRODUCTION

With non-face-to-face social interactions increasing in the wake of the COVID-19 pandemic, global over-the-top services have grown significantly [1]. Snack culture, in which short-format content is consumed, has simultaneously emerged as a new trend, and the proportion of short-form media used in images, videos, and news is also increasing [2]. Additionally, because media content is openly accessible, its influence on individuals and industries has increased.

YouTube, which is the most representative video platform, was launched in 2005. It distributes advertising revenue to publishers based on the number of subscribers and their views on their channels. According to a recent analysis of the US film industry, the more violent or provocative the content, the higher the profits in the international market [3]. Accordingly, video publishers create stimulating posts containing harmful content to increase interest and induce clicks on their videos, as part of efforts to increase revenue [4]. However, the primary consumers of the content in these videos are teenagers. In light of an increase in the Internet usage rate amid the COVID-19 pandemic, the frequency of exposure to harmful media has also soared; teenagers and young people are more vulnerable to such harmful content on social media [5,6].

According to a survey of American parents [7], 80% of children under the age of 11 watch videos on YouTube, with 85% of American boys and 70% of American girls aged 13-17 watching YouTube videos daily. Viewers can watch videos on YouTube without restrictions. Furthermore, unrestricted access to harmful and inappropriate content on YouTube videos is increasing among minors [7]. YouTube spends significant time to recognize harmful content, but its machine-learning algorithms that automatically identify problematic content find it challenging to recognize ambiguities, resulting in misclassification [8].

Received 4 January 2023, **Revised** 26 January 2023, **Accepted** 27 January 2023
***Corresponding Author** Il-Gu Lee (E-mail: iglee@sungshin.ac.kr, Tel: +82 2-920-7145)
Department of Convergence Security Engineering in College of Knowledge-based Service Engineering, Sungshin Women's University, 02844, Republic of Korea

Open Access https://doi.org/10.56977/jicce.2023.21.1.75 print ISSN: 2234-8255 online ISSN: 2234-8883

Panopto, a video content management software company, ranked YouTube as one of the best video platforms in 2023 [16]. YouTube has been at the forefront of video platforms since its launch in 2005 and it is now the most accessible and popular video platforms available. It is an open platform that can be viewed by all age groups and is also being used in public places, such as schools, to disseminate educational material. Accordingly, YouTube has introduced technology to prevent teenagers from engaging with harmful videos.

Violent media content has a significant impact on the psychology and perception of adolescents. Sexual images have a strong influence on the adoption of distorted sexual beliefs among adolescents [14]. Adolescents who are exposed to violent and suggestive media are approximately six times more likely to be aggressive and develop distorted views about sex than those with no such exposure. To address this problem, governments in developed countries have announced plans to strengthen their responses to the distribution of illegal and harmful digital materials. YouTube has also followed this trend with a separate setting known as "Restricted Mode" for restricting access to such videos.

Restricted Mode is a filtering function that filters videos after analyzing information such as metadata, title, and video description. In this mode, an automated system conducts the first review, followed by an administrator. The filtered video is periodically reviewed by the administrator to impose sanctions on content violations [9]. However, in this mode, users must activate the function directly to use it, and content creators can bypass filtering by mixing noise with thumbnails or inserting special characters between titles. Therefore, this mode has limitations in terms of usability and functionality [10]. Additionally, although various methods, such as extracting the hash value of a video or censoring uniform resource locators, are being studied to filter harmful media, they are not adequate for effectively responding to the detours that content creators take to avoid being penalized [11-13]. Therefore, it is necessary to study filtering technology that can prevent detours by using information that can be checked when a user clicks on a video. Since the conventional post-processing method based on user reports cannot properly manage the rapidly increasing and disseminated media, an automated method for quickly filtering media content is required.

Accordingly, this study proposes a filtering technique based on multimodal information extraction to compensate for the limitations of the existing filtering techniques. Filtering techniques based on the extraction of multiple pieces of information operate by excerpting text information from titles, thumbnails, and audio data, to perform forbidden word-based filtering. The contributions of this study are as follows:

- It is possible to detect harmful media quickly and accurately and respond to it using automated techniques to extract and filter information.
- Using a detection method based on a forbidden-word database, unreported media can be detected expeditiously and effectively than when using the conventional method, wherein actions are taken after reviewing the user report.

The remainder of this paper is organized as follows. Section II analyzes the existing filtering techniques, while Section III discusses the proposed filtering technique based on information extraction from multiple sources. Section IV evaluates and analyzes the proposed method by comparing it with the conventional method, and Section V proposes future research and concludes the study.

II. RELATED WORK

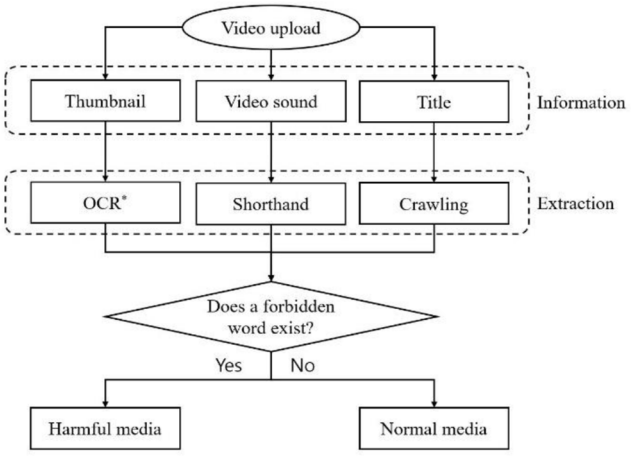
In this section, we introduce various techniques for detecting harmful videos and analyze those used to restrict harmful videos on YouTube.

A technique to detect whether images and videos are pornographic by determining the amount of skin tone was proposed in a previous study [15]. This method detects skin-color pixels in an image, derives a skin area based on the detected pixels, and then analyzes the skin area to determine whether the image contains nudity. Experiments were conducted with 986 harmful images and 253 harmful videos, and the technique achieved an accuracy rate of 80.23%. However, there was a limitation in that harmful images could not be accurately detected with only the percentage of skin color, and harmful words included in thumbnails or images could not be detected.

Another study proposed a deep-learning-based architecture to detect and classify inappropriate content [22]. Video descriptors and video representations were correspondingly extracted and learned using EfficientNet-B7, which is a transfer learning model based on convolutional neural networks. Once the learning was completed, multi-classification was performed based on the bidirectional long short-term memory network for effective learning. According to the experimental results, the EfficientNet bidirectional long short-term memory network model showed a high accuracy of 95.66%. However, since this study was a result of experiments wherein animations with relatively distinct colors or characteristics were used as learning data, there was a limitation in that violent or sensational acts or language-based media could not be detected without actions.

Previous studies have focused only on detecting harmful behaviors in the content of media items. Additionally, there is a common limitation among the methods in that they cannot detect sexual or violent content based on audio data in the media. Therefore, to solve this problem, the content of the video, which includes audio, needs to be filtered.

YouTube has also introduced and operated technologies to prevent teenagers from engaging in harmful media. The technique used by YouTube to restrict harmful videos generally involves two steps. First, the automated system built into YouTube checks the metadata, title, and language of the video, and restricts the video according to this information. Subsequently, if the user reports harmful videos that were not detected by the automated system, the person in charge directly checks the video and decides on actions such as disclosure, deletion, age restriction, or function restriction [17]. The Restricted Mode provided by YouTube restricts videos so that only public videos can be viewed [9]. This has a limitation in that videos containing queer-related content are misclassified as harmful videos that have not been reported, and they are mistakenly filtered out [18].



*OCR: Optical character recognition.

Fig. 1. Flowchart of the proposed technique.

III. IMPLEMENTATION OF HARMFUL-MEDIA FILTERING FRAMEWORK

In this section, we describe the proposed filtering technique, which is intended to overcome the limitations of existing methods, and is based on information that can be extracted from videos.

Therefore, the proposed technique can identify video types that cannot be detected using conventional filtering techniques. The details of these videos are discussed below:

Although the content is appropriate, the thumbnails or titles used to induce clicks are provocative.

Only the thumbnail and title are expressed as general titles to bypass the filtering of harmful videos.

The title, thumbnails, and content of a video are presented as new words, or detours are used to prevent the video from being filtered as harmful media.

By applying the proposed technique to a video before uploading it to the video platform, harmful media can be identified without employing a separate manual data-filtering process or user reports. Additionally, the proposed technique can filter not only searchable text, but also letters in thumbnails, making it possible to identify videos that induce clicks using provocative thumbnails and titles. Therefore, the limitations of the conventional method can be overcome using the proposed method.

A. Framework for Harmful-Media Filtering Technique

Fig. 1 shows the operating procedure of the proposed multi-filtering technique.

Multimodal filtering techniques include video information extraction, text conversion, and forbidden-word filtering. First, after extracting the thumbnail, audio, and title from the video, multiple pieces of information for each item are extracted, and the presence or absence of forbidden words is determined. The text in the thumbnail is extracted using optical character recognition (OCR) and the audio is converted into text information through a shorthand program. If forbidden words are included in these three pieces of text information, the content is deemed to be harmful.

For certain videos, harmful words are mixed in the thumbnails to induce clicks. Such videos are difficult to detect without writing titles and descriptions to bypass the detection algorithm. However, the proposed technique can also be used to detect such videos. OCR is a technique that recognizes and extracts characters in an image file. Currently, many open-source OCR techniques are open to the public for free, as they are provided in various services as application-programming interfaces. We applied the OCR technique using the open application-programming interface provided by Google [19].

When detecting a video that appears to be normal from its thumbnail and title, but contains harmful content, the audio information of the video is converted into text and used for detection. Voice shorthand is a recording method in which an individual's voice is written using a specific symbol and then converted into text. In this study, we used VREW [20], which is a machine-learning-based automatic shorthand program, to depict the audio in the video in shorthand.

Title filtering performs forbidden word filtering by crawling the titles in the video. We used the WebDriver command of Selenium, which is a Python open-source library for automatic website testing, and conducted the test by directly opening the ChromeDriver. Additionally, the title tag of the page was verified and extracted using the find command of Beautiful Soup, which is a library that aids in easy searching for information from HyperText Markup Language tags.

Voice filtering works by depicting the audio of the video in shorthand using the VREW program and applying forbid-

den-word filtering to the shorthand text file to determine the presence or absence of harmful words.

Forbidden-word filtering is a technique that compares the inputted text with words from a previously created database with harmful words. Forbidden words are collected mainly from words classified as profanities and harmful words in various communities; overall, 3,148 harmful words were collected [21]. Since the information collected for filtering is diverse and contains details such as titles, thumbnails, and audio in videos, it is important to determine a threshold value to optimize the filtering performance.

B. Multimodal-Information Extraction Techniques

We considered 11 filtering combinations, as listed in Table 1, to determine the conditions under which a video can be classified as harmful. Both cases wherein a single item and a combination of three items underwent filtering were considered. Regarding complex use, the most efficient method was sought by determining cases wherein all the items were classified as harmful (AND) and cases wherein even one of the items was classified as harmful (OR).

Table 1. Filtering combinations for multimodal-information extraction

Type		Filtering combination
Single		Title
		Thumbnail-OCR*
		Sound-Shorthand
Multi	Double	Title **AND** Thumbnail-OCR
		Title **AND** Sound-Shorthand
		Thumbnail-OCR **AND** Sound-Shorthand
		Title **OR** Thumbnail-OCR
		Title **OR** Sound-Shorthand
		Thumbnail-OCR **OR** Sound-Shorthand
	Triple	Title **AND** Thumbnail-OCR **AND** Sound-Shorthand
		Title **OR** Thumbnail-OCR **OR** Sound-Shorthand

*OCR: Optical character recognition.

IV. EVALUATION AND ANALYSIS

A. Data Environment Containing Known Videos

In this study, the Restricted Mode of YouTube and the proposed method were compared and evaluated in the same experimental environment. For the experiment, 100 videos that could be viewed without adult verification were selected and labeled after manually determining whether they were harmful to the accuracy of the experimental results. Among the 100 collected images, 53 were considered normal and 47 harmful.

Using the harmful media filtering framework, multimodal information in the video was collected, and forbidden-word filtering was applied. In the case of YouTube, which was the comparison target, the accuracy in the discrimination of harmful media was measured based on the results under the Restricted Mode.

Table 2 compares the performance of YouTube's Restricted Mode and the proposed technique. The Restricted Mode accurately classified 81 of the 100 videos as normal, and the proposed filtering technique classified 90 videos as normal. These results confirmed that the proposed technique performed 9% better than the conventional method.

Table 2. Comparison of performances of YouTube's Restricted Mode and proposed filtering technique

	Accuracy (%)	Precision (%)	Recall (%)
Proposed	90	100	78.7
Restricted Mode	81	75	89.3

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \tag{1}$$

$$\text{Precision} = \frac{TP}{TP+FP} \tag{2}$$

$$\text{Recall} = \frac{TP}{TP+FN} \tag{3}$$

Table 2 lists the accuracy, precision, and recall of both methods based on the experimental results. Accuracy indicates how accurately the technique predicts the harmful label in a video dataset, as represented in Eq. (1). Precision is the percentage of effectively harmful videos among those classified as harmful by the technique; it can be obtained using Eq. (2). Recall represents the number of harmful samples that are classified as harmful and is calculated using Eq. (3).

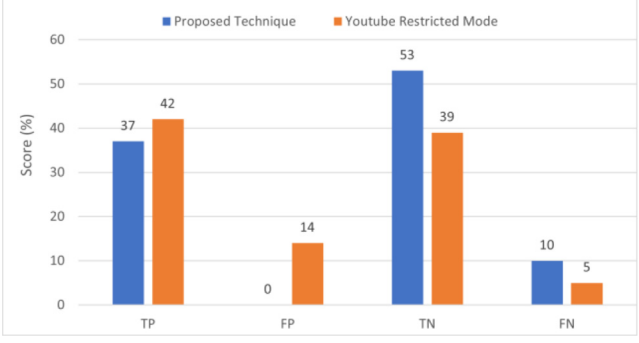


Fig. 2. Comparison of confusion matrices of proposed technique and YouTube's Restricted Mode in data environment with known videos.

Fig. 2 presents a comparison of the confusion matrices of the proposed technique and YouTube's Restricted Mode. Based on the experimental results, true positive (TP) indi-

cates that a harmful video has been correctly predicted as harmful, true negative (TN) indicates that a harmless video has been accurately predicted as harmless, false positive (FP) indicates that a normal video has been incorrectly determined as harmful (over-detection), and false negative (FN) indicates that a harmful video has been wrongly determined as normal (no detection). Fig. 2 shows that TPs and FPs occurred more frequently under YouTube's Restricted Mode, and TNs and FNs occurred more frequently under the proposed technique. In particular, YouTube's Restricted Mode makes more correct positive decisions than the proposed method.

B. Data Environment Containing Unknown Videos

YouTube's Restricted Mode is an automated system that manages harmful videos based on user reports. Therefore, to assess this system clearly, it is necessary to evaluate its performance in situations where user reports do not reflect in the results. Accordingly, new videos were produced to evaluate YouTube's automation system. The videos produced comprised 50 normal and harmful videos each. There were no forbidden words in the normal videos or their thumbnails and titles; however, forbidden words appeared more than once in the harmful videos. Thereafter, in an attempt to circumvent detection, 900 additional videos were created, wherein various noises were mixed with the thumbnails, videos, and audio of the produced videos. Practically, there may be cases wherein forbidden words are used unintentionally or for educational purposes, and such words may not exist in the forbidden word database that is used for filtering. To reflect such practical scenarios in the experiment, the above case was assumed, and 25 videos corresponding to each scenario were created and used in the experimental dataset. In the second experiment, all the 1,050-video data that were previously generated were used.

The second experiment was conducted in a similar manner, after uploading the produced video to YouTube and leaving it for 24 h. However, in this experiment, we attempted to determine the most optimal filtering technique by applying the 11 filtering combinations mentioned in Table 1, rather than all three methods: Title, Thumbnail-OCR, and Sound-Shorthand.

Table 3 presents the results of the second experiment using YouTube's Restricted Mode. Failure to detect the newly produced videos shown in Fig. 3 can be attributed to harmful videos being incorrectly classified as normal. In the first experiment, the exceptional performance of Restricted Mode was confirmed, as it accurately detected 42 of 47 harmful videos. However, in the second experiment, wherein user reports were not reflected in the results, YouTube's Restricted Mode detected only 96 of the 525 harmful videos as normal, suggesting that this method relies heavily on user reports.

Table 3. Comparison of performances of proposed method and YouTube's Restricted Mode (conventional method)

Filtering technique	Accuracy (%)	Precision (%)	Recall (%)
Title	88.57	94.5	81.9
Thumbnail-OCR*	88.57	94.5	81.9
Sound-Shorthand	68.57	89.79	41.9
Title **AND** Thumbnail-OCR	88.57	94.5	81.9
Title **OR** Thumbnail-OCR	88.57	94.5	81.9
Title **AND** Sound-Shorthand	62.47	86.18	29.71
Title **OR** Sound-Shorthand	**94.66**	**95.18**	**94.09**
Thumbnail-OCR **AND** Sound-Shorthand	62.47	86.18	29.71
Thumbnail-OCR **OR** Sound-Shorthand	**94.66**	**95.18**	**94.09**
Title **AND** Thumbnail-OCR **AND** Sound-Shorthand	62.47	86.18	29.71
Title **OR** Thumbnail-OCR **OR** Sound-Shorthand	**94.66**	**95.18**	**94.09**
YouTube restricted mode	53.71	62.74	18.28

*OCR: Optical character recognition

Table 3 presents the results of the second experiment, using the proposed method. Accuracy was the lowest when Sound-Shorthand was used alone. However, the best performance was achieved when using the OR method, which uses Title or Thumbnail-OCR, and classifies a video as harmful, even if only one of these two types of information is detected.

The three combinations showed correspondingly exceptional performances, but among them, Title OR Thumbnail-OCR OR Sound-Shorthand method was considered representative for the proposed method and selected for comparison with YouTube's Restricted Mode.

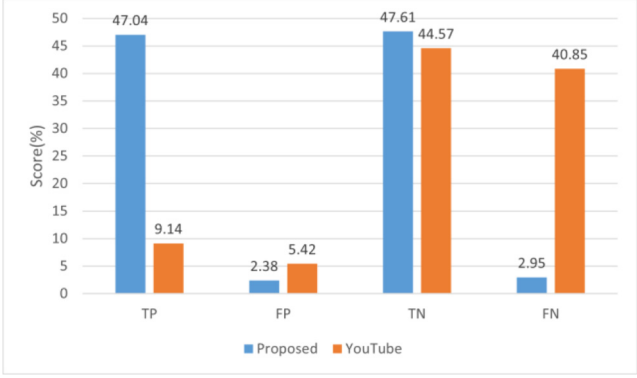


Fig. 3. Comparison of confusion matrices under proposed technique and YouTube's Restricted Mode in data environment with unknown videos.

Fig. 3 shows a comparison between the confusion matrices under YouTube's Restricted Mode and the proposed technique. Regarding TNs, both techniques performed similarly,

but in the case of TPs, the performance of YouTube's Restricted Mode was remarkably low. In addition, despite adding videos containing harmful words that were not present in the database and videos using homonyms of forbidden words to reflect possible practical situations, the false detection rate of Restricted Mode was approximately 3% higher than that of the proposed technique, while detection rate was approximately 38% higher. The positive decision rate was high under the Restricted Mode when user reports were included, but when these reports were not included, there were fewer TPs and FPs, and more TNs and FNs. Contrary to previous experimental results, the results obtained using YouTube's automated algorithm had a low rate of accurate decisions and a high rate of incorrect decisions.

The experimental results showed the high dependency of YouTube's Restricted Mode on user reports to provide good performance. To overcome these shortcomings, the proposed filtering technique based on multimodal information extraction was performed at the time of the video upload, and the contents and thumbnails of the video were automatically filtered, resulting in relatively fast and accurate results. Additionally, this approach is cost-effective because there is no need for separate personnel to check each reported video, as is the case in Restricted Mode.

V. CONCLUSION

This study proposed a novel filtering technique that uses a forbidden word-based filtering technique wherein three types of multimodal information, namely the thumbnail, title, and audio, from a video are used to filter harmful media. The text in the thumbnail was extracted through OCR, and the audio data were converted into a text file through an artificial intelligence shorthand program. Subsequently, forbidden word-based filtering was performed, and if one of the three types of information was considered forbidden, then the corresponding video was determined to be harmful. According to the experimental results, the proposed technique improved accuracy by 9% over that of YouTube's Restricted Mode and showed a 41% better detection accuracy than YouTube's Restricted Mode without the support of user reports.

The multimodal information-based filtering technique proposed in this study has certain limitations, such as reducing the number of FPs and FNs in the results, because it is a detection technique based on forbidden words. To overcome these limitations, in the future, we plan to study filtering techniques that can identify different contexts using artificial intelligence-based multimodal information.

ACKNOWLEDGMENTS

This work was supported by Sungshin Women's University Research Grant No. H20220029.

REFERENCES

[1] D. Madnani, S. Fernandes, and N. Madnani, "Analysing the impact of COVID-19 on over-the-top media platforms in India," *International Journal of Pervasive Computing and Communications*, vol. 16, no. 5, pp. 457-475, Aug. 2020. DOI: 10.1108/IJPCC-07-2020-0083.

[2] J. Nam and Y. Jung, "Digital natives' snack content consumption and their goals: A means-end chain approach," *Telematics and Informatics*, vol. 63, p. 101664, Oct. 2021. DOI: 10.1016/j.tele.2021.101664.

[3] S. Chukwu-Okoronkwo, D. Omeonu, and D. Onwuka, "Television and video films and the rhythm of violence: assessing the negative effect of youths' exposure to violent television and video films content," *New Media and Mass Communication*, vol. 92, pp. 15-24, Aug. 2020. DOI: 10.7176/NMMC/92-02.

[4] Z. Fan, K. Fan, Y. Tian, and W. Zhang, "Evolutionary game analysis of harmful information governance in social networks with public participation," in *Proceedings of the 2020 IEEE 3rd International Conference on Electronic Information and Communication Technology (ICEICT)*, Shenzhen, China, pp. 725-728, 2020. DOI: 10.1109/ICEICT51264.2020.9334207.

[5] E. Bozzola, G. Spina, R. Agostiniani, S. Barni, R. Russo, E. Scarpato, A. Di Mauro, A. V. Di Stefano, C. Caruso, G. Corsello, and A. Staiano, "The use of social media in children and adolescents: scoping review on the potential risks," *International Journal of Environmental Research and Public Health*, vol. 19, no. 16, p. 9960, Aug. 2022. DOI: 10.3390/ijerph19169960.

[6] M. Spitzer, "Open schools! Weighing the effects of viruses and lockdowns on children," *Trends in Neuroscience and Education*, vol. 22, p. 100151, Mar. 2021. DOI: 10.1016/j.tine.2021.100151.

[7] M. Hattingh, "The dark side of YouTube: a systematic review of literature," in *Adolescences*, London: InTechOpen, 2021.

[8] L. Kobilke and A. Markiewitz, "The Momo Challenge: measuring the extent to which YouTube portrays harmful and helpful depictions of a suicide game," *SN Social Sciences*, vol. 1, no. 4, pp. 1-30, Feb. 2021. DOI: 10.1007/s43545-021-00065-1.

[9] YouTube Help, Your YouTube content & Restricted Mode [Internet]. Available: https://support.google.com/youtube/answer/7354993?hl=en.

[10] R. Tahir, F. Ahmed, H. Saeed, S. Ali, F. Zaffar, and C. Wilson, "Bringing the kid back into YouTube Kids: Detecting inappropriate content on video streaming platforms," in *Proceedings of 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, Vancouver, Canada, pp. 464-469, 2019. DOI: 10.1145/3341161.3342913.

[11] F. Saurwein and C. Spencer-Smith, "Automated trouble: the role of algorithmic selection in harms on social media platforms," *Media and Communication*, vol. 9, no. 4, pp. 222-233, Nov. 2021. DOI: 110.17645/mac.v9i4.4062.

[12] D. Varshney and D. K. Vishwakarma, "Hoax news-inspector: a real-time prediction of fake news using content resemblance over web search results for authenticating the credibility of news articles,"

Journal of Ambient Intelligence and Humanized Computing, vol. 12, no. 9, pp. 8961-8974, Nov. 2020. DOI: 10.1007/s12652-020-02698-1.

[13] N. A. Khan, A. Khan, M. Ahmad, M. A. Shah, and G. Jeon, "URL filtering using big data analytics in 5G networks," *Computers & Electrical Engineering*, vol. 95, no. 3, p. 107379, Oct. 2021. DOI: 10.1016/j.compeleceng.2021.107379.

[14] E. W. Owens, R. J. Behun, J. C. Manning, and R. C. Reid, "The impact of Internet pornography on adolescents: A review of the research," *Sexual Addiction & Compulsivity*, vol. 19, no. (1-2), pp. 99-122, Jan. 2012. DOI: 10.1080/10720162.2012.660431.

[15] M. B. Garcia, T. F. Revano, B. G. M. Habal, J. O. Contreras, and J. B. R. Enriquez, "A pornographic image and video filtering application using optimized nudity recognition and detection algorithm," in *Proceedings of 2018 IEEE 10th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM)*, Baguio City, Philippines, pp. 1-5, 2018. DOI: 10.1109/HNICEM.2018.8666227.

[16] Panopto, Comparing the best video platforms in 2023. [Internet] Available: https://www.panopto.com/kr/blog/comparing-the-top-online-video-platforms/.

[17] YouTube Help, Report inappropriate content, channels, and other content on YouTube. [Internet] Available: https://support.google.com/youtube/answer/2802027.

[18] Google Cloud, Detect text in images. [Internet] Available: https://cloud.google.com/vision/docs/ocr?skip_cache=true.

[19] C. Southerton, D. Marshall, P. Aggleton, M. L. Rasmussen, and R. Cover, "Restricted modes: social media, content classification and LGBTQ sexual citizenship," *New Media & Society*, vol. 23, no. 5, pp. 920-938, DOI: 10.1177/1461444820904362.

[20] Vrew, Fast|Simple|Easy video editing. [Internet] Available: https://vrew.voyagerx.com/en/.

[21] Github, Bad-word-list-korean. [Internet] Available: https://github.com/YEON-JI-LEE/bad-word-list-korean.git.

[22] K. Yousaf and T. Nawaz, "A deep learning-based approach for inappropriate content detection and classification of YouTube videos," *IEEE Access*, vol. 10, pp. 16283-16298, 2022. DOI: 10.1109/ACCESS.2022.3147519.

Yeon-Ji Lee

received her B.S. degree in convergence security engineering from Sungshin Women's University in 2022, and she is currently pursuing an M.S. degree in future convergence technology engineering at the same university. Her current research interests include convergence security, malware detection, machine learning, and data analysis.

Ye-Sol Oh

received her B.S. degree in convergence security engineering from Sungshin Women's University in 2022, and she is currently pursuing an M.S. degree in future convergence technology engineering at the same university. Her current research interests include anomaly detection, digital forensics, and convergence security.

NA-Eun Park

received her B.S. degree in convergence security engineering from Sungshin Women's University in 2021, and she is currently pursuing an M.S. degree in future convergence technology engineering at the same university. Her current research interests include endpoint security, anomaly detection, system security, and artificial intelligence.

Il-Gu Lee

received his B.S. degree in electrical engineering from Sogang University, Seoul, Korea, in 2003, and his M.S. degree from the Department of Information and Communications Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea, in 2005. He received his Ph.D. degree from the KAIST Graduate School of Information Security in Computer Science & Engineering in 2016. He is a professor at the Department of Convergence Security Engineering and the Department of Future Convergence Technology Engineering, Sungshin Women's University (SWU), Seoul, Korea. Before joining SWU in March 2017, he worked at the Electronics and Telecommunications Research Institute (ETRI) as a senior researcher from 2005 to 2017, and served as a principal architect and project leader for Newratek (KR) and Newracom (US) from 2014 to 2017. His current research interests include wireless/mobile networks with an emphasis on information security, networks, wireless circuits, and systems. He has authored/coauthored more than 140 technical papers in the areas of information security, wireless networks, and communications, and holds approximately 160 patents. He is also an active participant in and contributor to the IEEE 802.11 WLAN standardization committee.