Computer Modeling in Engineering & Sciences

Tech Science Press

Doi:10.32604/cmes.2025.061653

ARTICLE

Deep Learning-Based Natural Language Processing Model and Optical Character Recognition for Detection of Online Grooming on Social Networking Services

Sangmin Kim[1], Byeongcheon Lee[1], Muazzam Maqsood[2], Jihoon Moon[3,*] and Seungmin Rho[4,*]

[1]Department of Security Convergence, Chung-Ang University, Seoul, 06974, Republic of Korea
[2]Department of Computer Science, COMSATS University Islamabad, Attock Campus, Attock, 43600, Pakistan
[3]Department of Data Science, Duksung Women's University, Seoul, 01369, Republic of Korea
[4]Department of Industrial Security, Chung-Ang University, Seoul, 06974, Republic of Korea

*Corresponding Authors: Jihoon Moon. Email: jmoon25@duksung.ac.kr; Seungmin Rho. Email: smrho@cau.ac.kr

Received: 29 November 2024; Accepted: 02 April 2025; Published: 30 May 2025

ABSTRACT: The increased accessibility of social networking services (SNSs) has facilitated communication and information sharing among users. However, it has also heightened concerns about digital safety, particularly for children and adolescents who are increasingly exposed to online grooming crimes. Early and accurate identification of grooming conversations is crucial in preventing long-term harm to victims. However, research on grooming detection in South Korea remains limited, as existing models trained primarily on English text and fail to reflect the unique linguistic features of SNS conversations, leading to inaccurate classifications. To address these issues, this study proposes a novel framework that integrates optical character recognition (OCR) technology with KcELECTRA, a deep learning-based natural language processing (NLP) model that shows excellent performance in processing the colloquial Korean language. In the proposed framework, the KcELECTRA model is fine-tuned by an extensive dataset, including Korean social media conversations, Korean ethical verification data from AI-Hub, and Korean hate speech data from HuggingFace, to enable more accurate classification of text extracted from social media conversation images. Experimental results show that the proposed framework achieves an accuracy of 0.953, outperforming existing transformer-based models. Furthermore, OCR technology shows high accuracy in extracting text from images, demonstrating that the proposed framework is effective for online grooming detection. The proposed framework is expected to contribute to the more accurate detection of grooming text and the prevention of grooming-related crimes.

KEYWORDS: Online grooming; KcELECTRA; natural language processing; optical character recognition; social networking service; text classification

1 Introduction

In modern society, amenities such as the Internet and social networking services (SNSs) allow people of all ages to easily obtain, share, and use each other's information. While this accessibility offers numerous benefits, it also leaves people vulnerable to threats such as voice phishing, stalking crimes, and digital sex crimes using personal information [1]. Due to the anonymity afforded by SNSs, digital sex crimes, such as online grooming, continue to increase, posing a significant threat to the safety of children and adolescents, who are the primary targets. Online grooming is the act of requesting inappropriate photographs and meetings after establishing a trusting relationship with the victim through an SNS or

chat app. As defined by Welner, grooming sex crimes against children and adolescents proceed in six stages [2]:

1. Target the victim;
2. Gain the victim's trust;
3. Fill a need;
4. Attempt sexual exploitation;
5. Maintain control;
6. Conceal the facts of the perpetration.

In the first stage, the perpetrator proceeds with a search to select victims. The perpetrator tends to consider the vulnerabilities and interests of children in the selection process. In the second stage, the perpetrator often pretends to be kind and interested in forming a relationship with the victim to gain trust. In the third stage, interactions with the victim are increased, and gifts or rewards are provided to satisfy the victim. In the fourth stage, the perpetrator makes a full-blown attempt at sexual exploitation. At this stage, the perpetrator checks the victim's response and will not relent if they reject their advances. In the fifth stage, the perpetrator tries to maintain physical and psychological control over the victim to continue the grooming relationship. In the final stage, when this relationship is discovered, the perpetrator may attempt to hide the offense and shift responsibility to the victim. Through these processes, perpetrators seek to exploit vulnerabilities in victims, including familial, physical, economic, and psychological problems, to engage in sexual exploitation, exert control over them, and manipulate them. Fig. 1 provides a summary of the stages of online grooming.

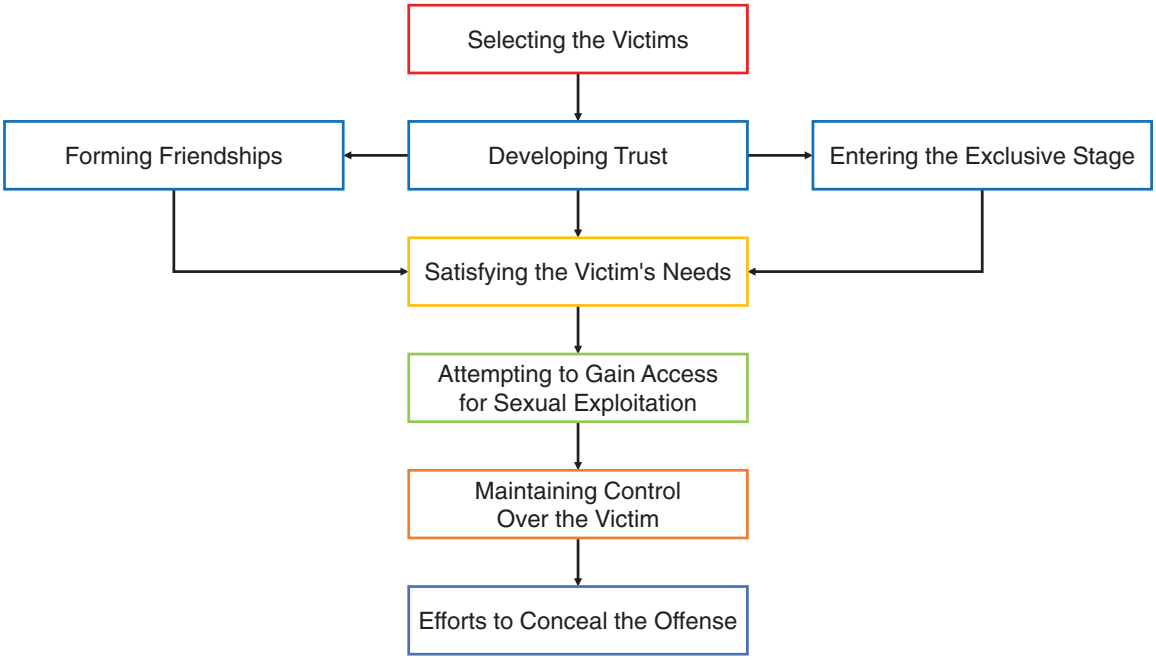


Figure 1: Summary of online grooming stages

The increasing anonymity attainable on the internet enables perpetrators of online grooming crimes to target children and adolescents easily and indiscriminately in cyberspace for sexual purposes [3]. This phenomenon can escalate into a significant societal problem due to the growing rate of sexual grooming crimes

through online channels. A survey conducted among 3789 adolescents aged 12–19 years in South Korea revealed that approximately 10% of them reported having met strangers online and in person. Additionally, another 10% from the same group reported having received gifts from strangers online [4]. These findings indicate that online grooming constitutes a severe problem affecting adolescents. Furthermore, South Korea recently witnessed a major criminal case known as the "Nth Room Case" [5]. This case involved crimes such as exploiting minors online and sharing sexually exploitative material obtained during the process through the Telegram mobile messaging service. This has significantly raised societal awareness regarding online grooming, and South Korea is currently exploring measures to address and prevent this problem.

To combat the online grooming of children and adolescents, South Korea introduced special investigative provisions for undercover operations targeting sexual crimes against minors in 2021 [6]. However, two years after the implementation of these regulations, concerns were raised regarding the effectiveness of undercover investigations. To apprehend perpetrators, police officers have participated in undercover operations, concealing their identities and engaging in open chat rooms and random chat applications. However, being adults, undercover officers face challenges in mimicking the behavior of actual children or adolescents, which may alert the perpetrators about the undercover operation being conducted; this complicates the enforcement of the new regulations. This leads to unnecessary resource allocation and is inefficient, as only seven individuals have been apprehended over approximately two years. Table 1 presents the status of police investigations of digital sexual crimes through undercover operations from September 2021 to September 2023. Given this scenario, the integration of artificial intelligence (AI) technology into such investigations is deemed essential. By learning the patterns in grooming conversations and conducting analyses to detect the conversation content, AI technology can alleviate the labor-intensive burden of undercover operations, providing more effective and efficient strategies to combat online grooming crimes [7].

Table 1: Police arrests from undercover investigations into digital sex crimes over three years

Charge	Identity private	Undercover	Total
Sale and distribution of material depicting sexual exploitation of adolescents	530	68	598 (41 arrests)
Possession and viewing of material depicting sexual exploitation of adolescents	–	118	118 (3 arrests)
Distribution of illegal videos	14	7	21 (1 arrest)
Online grooming	–	7	7 (4 arrests)

Although online grooming crimes are becoming increasingly severe, research on early prevention and detection of grooming conversations remains minimal. Existing solutions to prevent online grooming rely on simplistic approaches, such as blocking harmful websites, activity monitoring to display records of actions, or parental supervision and education [8]. However, these methods are largely ineffective in preventing online grooming, which highlights the need for more sophisticated technological solutions. Existing AI models are primarily trained on English text or fail to reflect the characteristics of SNS conversations, posing challenges in text classification. The Korean language presents even greater difficulties due to its structural and grammatical differences from English. Therefore, to improve the processing of Korean data, selecting and training models tailored to the characteristics of the Korean language is crucial.

To detect online grooming in SNS environments, this paper proposes a combination of deep learning-based natural language processing (NLP) models specialized for spoken Korean, such as KcELECTRA,

short for Korean comments ELECTRA (Efficiently Learning an Encoder that Classifies Token Replacements Accurately), and optical character recognition (OCR). This study is the first to propose a direct approach to detecting online grooming in Korean by integrating both OCR and NLP techniques. We construct text classification models based on this framework and evaluate them on new sentences. We assess the performance of the proposed framework by applying OCR algorithms to extract text from images and integrating these algorithms into the model. To validate the generalizability of the proposed framework, we compare their performance with transformer-based language models commonly used in many studies. Therefore, this study aims to leverage deep learning-based OCR and NLP to detect and block harmful texts across datasets, creating a safe environment for adolescents to engage in social networking. The primary contributions of this paper are as follows:

1. Innovative computational techniques: This paper introduces KcELECTRA, a pioneering NLP model specifically tailored to the analysis of colloquial Korean. Its development marks a significant advance in computational linguistics and sets a new benchmark for language models focused on non-English languages. This contribution is critical given the global nature of Internet use and the need for technologies that can operate effectively across diverse linguistic landscapes.
2. Enhanced recognition capabilities for digital communications: By integrating OCR with our NLP model, this study bridges the gap between textual analysis and visual data interpretation, improving the ability to detect grooming behaviors embedded in image-based social media interactions. This methodological innovation not only pushes the boundaries of digital communication monitoring but also introduces a novel approach to protecting minors in increasingly multimedia online environments.
3. Sociological insights into online interactions: From a sociological perspective, this research provides profound insights into the dynamics of online grooming, highlighting how linguistic nuances and cultural contexts influence predatory behaviors. By preserving the integrity of emotional expression and contextual cues in the analysis of SNS data, this study contributes to a deeper understanding of the social mechanisms underlying online exploitation, providing valuable insights for policymakers and educators in developing more effective prevention strategies.
4. Practical implications for online safety: The operationalization of this research provides practical tools for real-time monitoring and prevention of online grooming, significantly improving the safety protocols of social networking sites. This is particularly important in the digital age, where proactive measures are essential to protect vulnerable populations. The development of such tools not only demonstrates the applicability of advanced computational models in everyday technology but also highlights the role of academic research in driving technological solutions to societal problems.

The remainder of this paper is organized as follows: Section 2 discusses related work to provide context for our research. Section 3 describes the dataset and preprocessing methods as well as the development of the model and the OCR techniques used. Section 4 details the outcomes of our research, offering a comprehensive analysis of the models' performance, and discusses the immediate implications of our findings within the current research landscape. The paper concludes with Section 5, which summarizes our findings and highlights the implications and scope for future work.

2 Related Work

Research related to online grooming has been conducted in South Korea and many other countries. However, compared with research cases abroad, studies on online grooming crimes targeting children and adolescents in South Korea are severely lacking [3,9]. Moreover, even in the existing research, the focus has primarily been on comparing and analyzing legal systems or proposing response strategies through empirical analyses of online grooming. Because discourse on technological approaches to apprehend

perpetrators effectively is almost nonexistent, such research is essential for the protection of children and adolescents. Table 2 summarizes recent research on online grooming.

Table 2: Summary of recent studies on online grooming

Author (Year)	Research objective	Dataset	Method
Hong et al. (2023) [1]	To present criteria for legal intervention from the moment the sexual conversation is attempted	Random chat application conversations	LDA[a]
Milon-Flores et al. (2022) [8]	To detect grooming early using behavioral features	Schema-Guided Dialogue Dataset & PJZ and PJZC Datasets	BF-PSR[b]
Anderson et al. (2019) [10]	To identify child grooming in online chat conversations	200 text records & 1000 XML records	CDFTSVM[c]
Bours et al. (2019) [11]	To detect sexual predators in online chat conversations	PAN-2012 Perverted Justice	Naïve Bayes with TF-IDF[d]
Guo et al. (2023) [12]	To proactively prevent cybergrooming and protect youth from potential harm	Perverted Justice ConvAI dataset	T5[e]

Note: [a]latent Dirichlet allocation; [b]behavioral feature profile-specific representations; [c]coordinate-descent fuzzy-twin support vector machine; [d]term frequency-inverse document frequency; [e]text-to-text transfer transformer.

A study conducted in South Korea analyzed conversations from random chatting applications [1], employing topic modeling (latent Dirichlet allocation, LDA) to identify grooming types by topic. The harmonic mean was calculated to determine the K value and thus the optimal number of topics. Then, LDA was conducted by gradually increasing the K value, and the model was built using the Gibbs sampling method. The study categorized frequently occurring online grooming conversation topics into exploratory types: grooming Type 1, grooming Type 2, and high-risk. While this method captures the patterns and characteristics of online grooming conversations and could assist in revisions of online grooming punishment laws and legal standards, it does not involve practical detection through conversation.

Milon-Flores et al. [8] analyzed methods for early grooming detection in online conversations using the Schema-Guided Dialogue dataset and other resources. They developed a framework called behavioral feature (BF) profile-specific representation (PSR), combining seven additional BFs with an existing PSR. The PSR was modified through the model development process. The original PSR used filtered data, whereas the modified PSR employed the term frequency-inverse document frequency and maintained the writing styles of users by preserving tokens as unigrams without removing stopwords. This enhancement improved model performance. Combining BF-PSR with a multilayer perceptron yielded high accuracy, even with large imbalanced datasets. Furthermore, benchmark experiments demonstrated that behavioral features contribute to the early detection of online grooming.

Anderson et al. [10] developed an automated system for detecting grooming conversations using 200 text files related to child grooming and nongrooming conversations and 1000 XML files. They proposed a coordinate-descent fuzzy-twin support vector machine model that employs a radial basis function kernel for online grooming text classification. The model removes noise and reduces computational complexity using the coordinate descent algorithm, achieving improved classification with an accuracy of 60.96%. The paper

emphasizes that automated systems for grooming detection are becoming increasingly crucial for analyzing text in online conversations and suggests that integrating AI technology can enhance such systems.

Bours et al. [11] combined three approaches (message-based, author-based, and conversation-based) with five classification algorithms and two feature sets to detect perpetrators in online chat conversations. The paper emphasizes the importance of early detection, as perpetrators can cause harm before online conversations are analyzed. To determine the number of conversation messages needed to detect perpetrators, they used the naïve Bayes method. Testing with term frequency-inverse document frequency features resulted in a recall of over 0.8 after analyzing 36 messages using the five classifiers. This finding confirms the feasibility of early detection of online grooming, which necessitates further research on the possibility of using warning systems to alert users to suspicious grooming conversations.

Guo et al. [12] proposed a novel approach to cybergrooming prevention, noting a significant lack of research in understanding the characteristics of potential victims in grooming scenarios. This paper introduced a pioneering generative chatbot framework called Stop cybERgroomIng (SERI) to address the aforementioned gap. The SERI framework facilitates realistic conversations between perpetrators and chatbots posing as potential victims, contributing to the prevention of cybergrooming and the protection of adolescents. This innovative framework advances the development of preventive mechanisms and enhances the safety and awareness of vulnerable adolescents in online spaces.

In conclusion, research on early detection of online grooming is active outside South Korea, with behavioral characteristics being utilized and generative chatbot frameworks being developed to address regulatory problems and maximize cybergrooming awareness to protect adolescents. However, such research is lacking in South Korea. Therefore, we propose a comprehensive approach to address South Korea's online grooming problem by combining NLP and OCR. NLP facilitates the analysis of grooming conversations, whereas OCR extracts text from images of SNS conversations, enabling the detection of suspicious grooming dialogue. This approach surpasses the limitations of existing methods, which primarily involve blocking harmful sites or manual monitoring. The proposed solution holds promise in fostering a safer SNS chatting environment for children and adolescents, thereby alleviating a key societal concern.

3 Materials and Methods

Fig. 2 outlines the overall framework of our study. This section provides an overview of the datasets obtained from AI-Hub, the preprocessing steps applied to textual data and digital images, and the methods used to train and evaluate text classification models such as RoBERTa, DistilBERT, KoBERT, BLOOMz, DeBERTa-v3, FNet, and KcELECTRA. The section details how data are prepared and processed to support the detection of various types of online grooming behavior using advanced NLP techniques.

3.1 Data Preprocessing

This section describes the dataset used to build the text classification model and the preprocessing steps applied to this dataset. Additionally, it explains the preprocessing steps performed on the grooming conversation images to apply the extracted text to the constructed model.

We used Korean SNS data and text ethics verification data provided by AI-Hub [13,14], along with Korean hate speech data and images of online grooming conversations. The Korean SNS dataset consists of 2 million daily conversation records, categorized into nine topics (e.g., leisure, personal and interpersonal relationships, beauty, and health) in JavaScript Object Notation (JSON) format. This dataset primarily comprises casual conversations between men and women, with personal information anonymized and metadata (e.g., speaker information, conversation type, and topic) included.

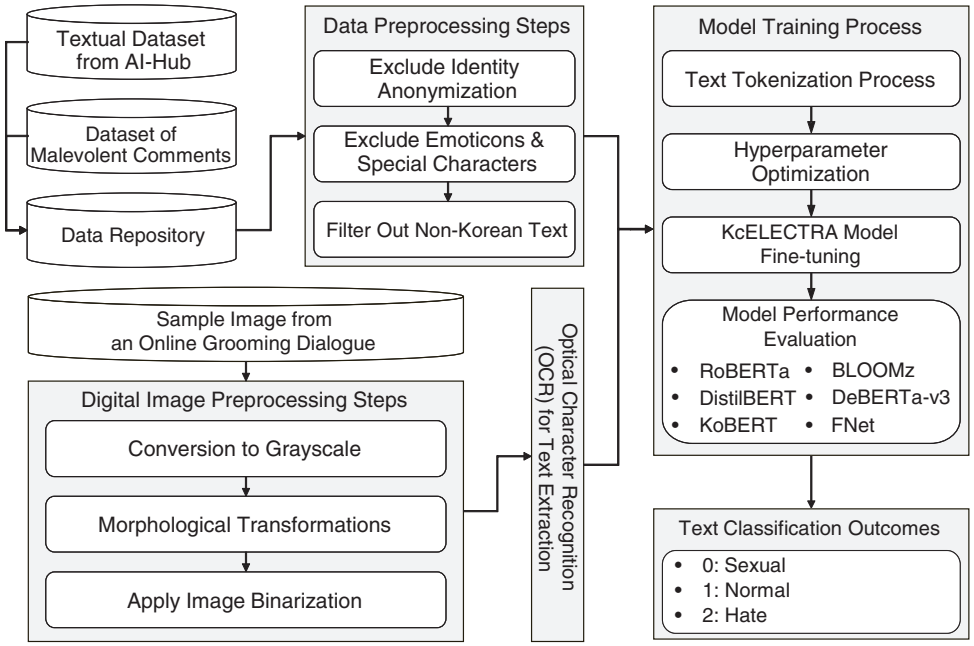


Figure 2: Overall research framework

The text ethics verification dataset contains 250,000 unethical sentences, classified into seven categories (e.g., abuse, violence, and sexual content). It was sourced from online community comments, SNS chats, and chatbot interactions. Each conversation set averages three sentences, with names of individuals or products anonymized to prevent social controversy. For this study, we specifically selected sentences from the sexual content category to focus on detecting grooming-related language.

To detect threats and hateful expressions that perpetrators may use to control victims in grooming conversations, we also incorporated Korean hate speech data [15]. This dataset includes approximately 100,000 utterances extracted from Korean online news comments between January 2018 and June 2020, classified into eight detailed categories (e.g., politics and appearance). Finally, the developed text classification model was tested with sample images of grooming conversations collected from SNSs using Google search. Table 3 provides a summary of the dataset statistics, and Fig. 3 illustrates examples of the collected text data.

Table 3: Statistical data summary

Dataset	Content	Type	Size
Korean SNS dataset	Speaker information, conversation type, and topic	JavaScript object notation	2,000,000 utterances
Korean text ethics verification dataset	Sentence type and speaker information	JavaScript object notation	251,064 utterances
K-MHaS dataset	Labeled with eight fine-grained hate speech classes	Plain text	109,062 utterances
SNS online grooming conversation images	Content of conversation	JPG and PNG	100 images

No.	Korean Conversation	English Translation
1	응 자연스러운게 나은거같아	Yeah, going natural seems better.
2	29일에 마지막 토익시험이거든	Got my final TOEIC test on the 29th.
3	ㅠㅠㅠ 여행가고싶닼ㅋㅋㅋ	Ugh, I'm dying to go on a trip lol
4	더러운 종교집단이다	That's a damn cult, seriously.
5	존나 아는척하는거 진짜 ㅈ같네	God, I can't stand know-it-alls.
6	그냥 시키는 대로 하지 반항을 하네	Why can't you just follow orders instead of rebelling?
7	점심부터 굶을까 생각중이야	Thinking of just skipping lunch from now on.
8	브라 입고 찍은 거 드릴게요	I'll send you the one where I'm wearing a bra.
9	알몸으로 찍어야 하나요?	Should I take one naked?
10	그럼 너 옷벗고 사진 좀 찍어서 보내봐	Then strip and shoot me a pic.

Figure 3: Examples of text data

To further validate the model and assess the potential for overfitting, we collected an additional set of 100 sentences for each category: normal, sexual, and hate sentences for evaluation. The normal sentence data were obtained from question-and-answer style chat logs within Korea's leading messaging application, KakaoTalk [16]. The sexual sentence data were translated into Korean from grooming sentences found in the PJZC dataset [17], a widely used international resource for online grooming detection. This dataset contained conversation data labeled as either non-groomer or groomer. The hate sentence data were obtained from the Korean hate speech dataset available on Kaggle [18], which was collected from Discord chat data. The data were provided in CSV, TXT, and TSV formats, respectively, in order to comprehensively evaluate the model's performance across different sentence structures and formats.

Preprocessing is necessary to perform text analysis on collected data. Typically, preprocessing of text data involves the removal of stop words. In this study, however, the removal of stop words such as profanity and slang were minimized in order to preserve the linguistic characteristics of the speaker's utterances in SNS conversations. This approach ensures that critical words conveying emotion or intent are not lost, allowing the model to accurately understand the context of the original utterances. Preprocessing includes the removal of personal information such as account numbers and phone numbers that have been anonymized (e.g., #@name#, #@number#). Emoticons, special characters, and unnecessary Japanese or English text were also removed. In addition, sentences that were ambiguous or difficult to interpret and consisted of three words or less (e.g., exclamations or simple answers to questions) and longer utterances consisting only of consonants (e.g., lol, haha) were removed. These steps help to improve the quality of the text data and the performance of the model. The preprocessed data were labeled as follows: sexually explicit: 0, normal speech: 1, and hate speech: 2. Table 4 summarizes the composition of the datasets after preprocessing, categorized by sentence type.

The image data collected at various resolutions ranging from low to high quality for text extraction using OCR algorithms undergoes preprocessing in the following five steps [19]:

1. Convert to grayscale;
2. Perform morphology operations;
3. Extract prominent edge boundaries;

4. Perform closing operations;
5. Define regions of interest.

Table 4: Construction of datasets by sentence type

Sentence type	Dataset			
	Training	**Validation**	**Testing**	**Total**
Normal	24,182	11,387	3981	40,000
Hate	18,037	9078	2945	30,060
Sexual	13,438	6914	2350	22,702

In Step 1, the original image was converted to grayscale to simplify pixel intensity representation, reducing computational complexity while preserving essential structural information. This transformation enhances the contrast between text and background, facilitating subsequent processing stages by improving text region distinction. In Step 2, morphological operations employing the gradient method were utilized to enhance edge detection within the image. This method calculates the difference between dilation and erosion operations, where dilation enlarges the boundaries of foreground objects while erosion reduces them. By emphasizing local intensity variations, the gradient method effectively sharpens text contours and speech bubble boundaries, facilitating more precise segmentation. This process enhances text distinguishability, particularly in SNS chat images, making text extraction more effective. In Step 3, the Otsu thresholding algorithm was employed to segment the extracted edges into a binary representation. By minimizing intra-class variance, this method dynamically determines the optimal threshold, enabling effective differentiation between text and background. In Step 4, closing operations were performed by applying dilation followed by erosion to the extracted edge boundaries. This process facilitates the connection of disjointed text components, bridging small gaps between characters and ensuring that fragmented textual elements are recognized as cohesive units. Given the prevalence of overlapping and handwritten text in SNS conversations, this operation substantially enhances text connectivity, thereby improving extraction accuracy. In Step 5, we defined regions of interest by clustering high-density text areas identified in previous steps. This process enabled the precise localization of text within the image, allowing for the final extraction of relevant text elements while filtering out unnecessary background components.

Fig. 4 illustrates the image preprocessing and shows the results obtained after preprocessing a low-resolution original image (643 × 907). Despite the challenges associated with reduced resolution, our OCR system effectively recognized the text in the message boxes, demonstrating a promising level of resilience. This preprocessing example highlights the ability of our model to maintain performance under less-than-ideal conditions by adaptively handling lower-quality images. Similarly, Fig. 5 shows the OCR preprocessing and results for a high-resolution image (1080 × 2058). In this case, our system performs consistently well, accurately recognizing text and demonstrating its robustness over a range of input data qualities. These results confirm the model's ability to handle variations in image resolution without significant loss of reliability or performance.

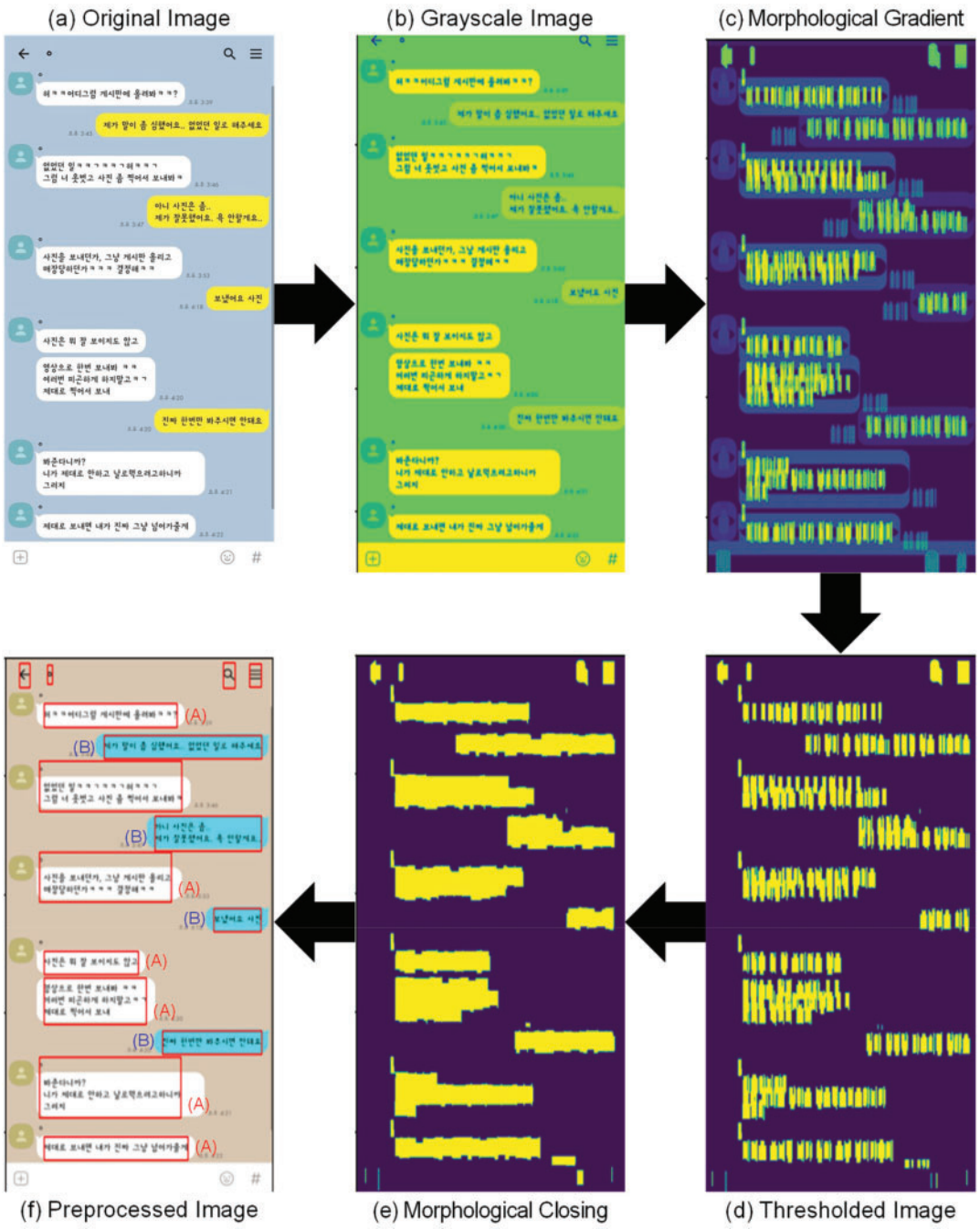


Figure 4: OCR preprocessing and result for low-resolution image (643 × 907). (a) Original Image: Low-resolution SNS chat image before preprocessing; (b) Step 1: Image converted to grayscale to simplify pixel intensities and enhance text-background contrast; (c) Step 2: Edges enhanced using the gradient method (difference between dilation and erosion) to sharpen text and speech bubble boundaries; (d) Step 3: Binary segmentation applied using Otsu's method to distinguish text from the background by minimizing intra-class variance; (e) Step 4: Dilation followed by erosion connects fragmented text components, bridging gaps to improve accuracy; (f) Step 5: Text regions localized by clustering high-density areas for precise extraction, filtering out irrelevant components

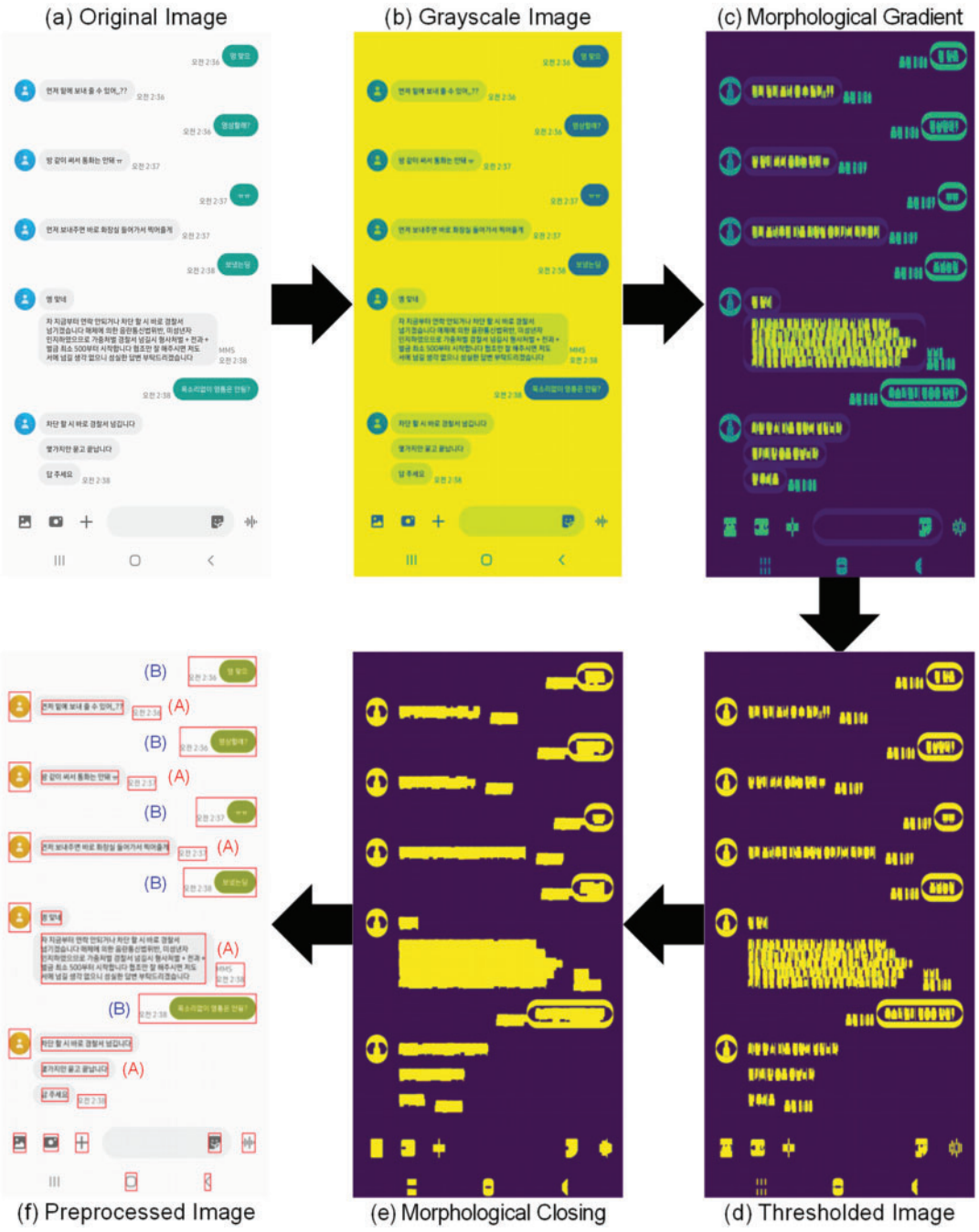


Figure 5: OCR preprocessing and result for high-resolution image (1080 × 2058). (a) Original Image: High-resolution SNS chat image prior to preprocessing; (b) Step 1: Simplifies the original image to grayscale, enhancing the distinction between text and background; (c) Step 2: Highlights edges of text and speech bubbles using morphological gradient methods, improving boundary clarity; (d) Step 3: Converts the image to a binary format by automatically determining the optimal threshold, clearly separating text from background; (e) Step 4: Connects fragmented text elements using dilation followed by erosion to bridge small gaps and enhance text continuity; (f) Step 5: Identifies and clusters high-density text regions for precise extraction, removing unnecessary background areas

By providing concrete examples of the OCR component's performance at different image resolutions, these figures support our claims about the adaptability and robustness of our OCR technology. This demonstrates our confidence in the effectiveness of OCR in different digital environments and supports the overall reliability of our model, even in the face of varying data quality. In addition, Fig. 6 shows the translation of a Korean conversation between individuals A and B into English, as shown in the preprocessed image. This translation helps provide context for the content processed by our OCR system to ensure clarity for readers unfamiliar with Korean.

Speaker	Korean Conversation	English Translation
A	허ㅋㅋ어디그럼 게시판에 올려봐ㅋㅋ?	Haha, go ahead and post it on the forum then, why don't ya? LOL
B	제가 말이 좀 심했어요.. 없었던 일로 해주세요	I spoke too harshly... Please, can we pretend it didn't happen?
A	없었던 일ㅋㅋㄱㅋㅋㄱ허ㅋㅋㄱ 그럼 너 옷벗고 사진 좀 찍어서 보내봐ㅋ	Pretend it didn't happen, LOL. Now, strip and send me a pic, come on.
B	아니 사진은 좀.. 제가 잘못했어요 욕 안할게요..	Not the picture, please...
A	사진을 보내던가, 그냥 게시판 올리고 매장당하던가ㅋㅋㅋ 결정해ㅋㅋ	Send the picture, or just post it yourself and get buried in shame LOL. Make up your mind.
B	보냈어요 사진	I sent the photo.
A	사진은 뭐 잘 보이지도 않고	I couldn't see picture well.
A	영상으로 한번 보내봐ㅋㅋ 여러 번 피곤하게 하지말고ㅋㄱ 제대로 찍어서 보내	The photo doesn't show much. Send a video, and don't make me ask again LOL. Do it right.
B	진짜 한번만 봐주시면 안돼요?	Can't you just let it go this one time, please?
A	봐준다니깐? 니가 제대로 안하고 날로먹으려고하니깐 그러지	I said I would look at it, didn't I? It's because you're trying to get off easy.
A	제대로 보내면 내가 진짜 그냥 넘어가줄게	Send it properly, and I'll really let it slide this time.

Figure 6: Conversion of Korean conversation to English in preprocessing

3.2 Model Development

This section explains the models and algorithms. The ELECTRA model comprises two network structures: a generator and a discriminator. The generator employs a pretraining method similar to that of the bidirectional encoder representations from transformers (BERT), using only the masked language model without the next sentence prediction (NSP) process. Given an input token sequence x, approximately 15% of tokens are masked to create a set of masked positions (Eq. (1)). The generator outputs the probability of generating token x_t at a given position t for the masked token X^{masked} (Eq. (2)). Therefore, the generator predicts and generates the original tokens, referred to as $X^{corrupted}$, which are passed to the discriminator (Eq. (3)). The discriminator performs binary classification on each token in the sequence generated by the generator, determining whether each token is original or replaced (Eq. (4)). The following are the formulas used in the generator and discriminator operations:

$$X^{masked} = \text{REPLACE}(x, m, [\text{MASK}]), \tag{1}$$

$$PG(x_t|x) = \exp(e(x_t)^T h_G(x)_t)/\text{sum}\{x`\}exp(e(x`)^T h_G(x)_t), \tag{2}$$

$$X^{\text{corrupted}} = \text{REPLACE}(x, m, \hat{x}), \tag{3}$$

$$D(x,t) = \text{sigmoid}(w^T h_D(x)_t). \tag{4}$$

Thus, ELECTRA is implemented using the generator to replace some tokens in the actual input with plausible fake tokens, and the discriminator predicts whether each token is real (from the actual input) or a generated fake token. A new pretraining task called "replaced-token detection" is performed to train the model on all input positions simultaneously; by allowing training on all tokens, this enables more efficient and rapid training compared with BERT, overcoming the 15% training limit on the input in the conventional masked language model (MLM). Fig. 7 illustrates the replaced-token detection (RTD) structure of the ELECTRA model.

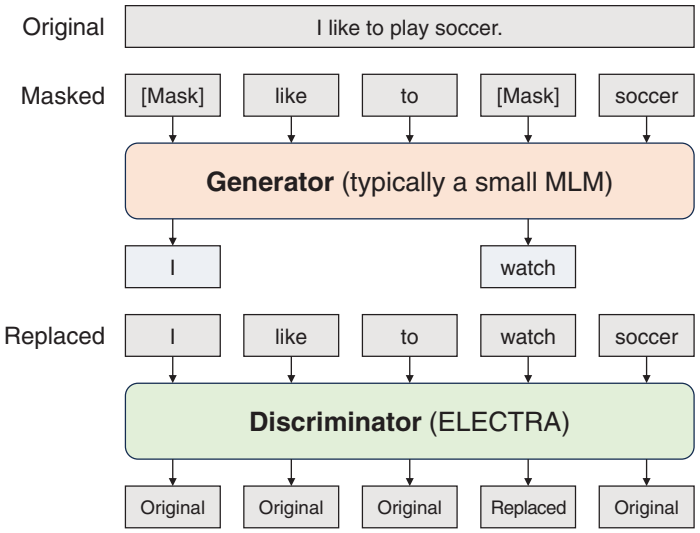


Figure 7: Overview of replaced-token detection

In this study, the KcELECTRA-base-v2022 model developed by Lee was employed for Korean text classification [20]. Recent transformer-based models for Korean text classification have primarily been trained on well-curated data, such as Korean Wikipedia, news articles, and books. In contrast, datasets such as the Naver Sentiment Movie Corpus are unrefined, containing colloquialisms, slang, and typographical errors not commonly found in formal writing. To account for such characteristics, KcELECTRA was pre-trained on comments and replies from Korean news articles, demonstrating excellent performance on user-generated and noisy text. Table 5 presents the performance of KcELECTRA on various datasets [20].

Table 5: Performance of KcELECTRA on various datasets

Model	Size	Datasets			
		Naver sentiment movie corpus (acc) [21]	NaverNER (F1) [22]	PAWS[a] (acc) [23]	KorNLI (acc) [24]
KcELECTRA-base-v2022	475 M	91.97	87.35	76.50	82.12
KcBERT-Base	417 M	89.62	84.34	66.95	74.85

(Continued)

Table 5 (continued)

Model	Size	Datasets			
		Naver sentiment movie corpus (acc) [21]	NaverNER (F1) [22]	PAWS[a] (acc) [23]	KorNLI (acc) [24]
KoBERT	351 M	89.63	86.11	80.65	79.00
XLB-RoBERTa-Base	1.03 G	89.49	86.26	82.95	79.92

Note: [a]Paraphrase adversaries from word scrambling.

3.3 Proposed Framework Process

The proposed framework for detecting online grooming behaviors follows a structured processing pipeline that integrates OCR and a fine-tuned KcELECTRA-based classification model. This framework systematically processes SNS chat images to extract and classify text, enhancing the accuracy of detecting harmful conversations. The overall workflow of the proposed framework is as follows:

1. Text data preprocessing: In this step, anonymized text data provided by AI-Hub undergoes preprocessing, where emoticons and special characters are removed, and non-Korean text is filtered out to retain only relevant linguistic content. Additionally, sentences that are ambiguous or difficult to interpret—such as those with three or fewer words or those consisting only of consonants—are excluded.
2. Image data preprocessing: SNS chat images, including screenshots from online platforms, contain user-generated conversations with varying text placements, fonts, and resolutions, making direct analysis challenging. To enhance text readability and optimize OCR performance, the input image undergoes preprocessing, which includes grayscale conversion, morphological operations, edge detection, closing operations, and region of interest (ROI) selection. These steps improve text visibility and remove background noise.
3. OCR-based text extraction: This step extracts the text embedded in the preprocessed image using OCR technology. The enhanced text visibility and reduced background noise from the previous step contribute to more accurate extraction.
4. Sentence segmentation and tokenization: In this step, the extracted text undergoes sentence segmentation and tokenization to maintain conversational context while preparing it for model input. Tokenization structures the data, enabling the deep learning model to process each linguistic unit effectively.
5. Model fine-tuning: The KcELECTRA model, optimized for Korean text classification, is fine-tuned using the tokenized text data. Trained specifically on informal, user-generated text, this model effectively distinguishes colloquial expressions, slang, and variations in writing styles commonly found in SNS conversations.
6. Text classification: Finally, the fine-tuned KcELECTRA model classifies each sentence into one of three categories:
   - Neutral speech: General conversations without harmful intent.
   - Sexually explicit speech: Sentences containing sexual language or solicitation.
   - Hate speech: Statements expressing discrimination or verbal abuse.

With this structured and systematic approach, the proposed framework, which seamlessly integrates OCR-based text extraction with deep learning-based classification, significantly enhances the accuracy of detecting online grooming and harmful conversations.

4 Results

This section explains the evaluation and performance metrics used in this study, along with the results obtained using the model on the collected dataset and the text extracted through OCR. Descriptions of the three benchmark models are provided, along with a comparative analysis of the performance of all models. The evaluations were performed on a computing environment consisting of a Windows 11 operating system, an Intel i7-12700K (12th generation) CPU, and an NVIDIA GeForce RTX 4080 (16 GB) GPU, supplemented by 64 GB of RAM. The system runs Python version 3.9. This hardware configuration, while robust, is typical of advanced consumer-level computing setups and not indicative of specialized high-performance computing facilities.

4.1 Performance Metrics

The detection of online grooming-related text is regarded as a multiclass classification problem involving suspicious grooming sentences, nongrooming sentences, and hate expressions. Standard performance metrics, namely accuracy (*Acc*), precision (*Pr*), recall (*Re*), and F1-score (the harmonic mean between precision and recall), were employed in this study:

$$Acc = (TP + TN)/(TP + FP + TN + FN), \tag{5}$$

$$Pr = TP/(TP + FN), \tag{6}$$

$$Re = TP/(TP + FN), \tag{7}$$

$$F1 - score = 2 \times (Pr \times Re)/(Pr + Re), \tag{8}$$

where true positive (*TP*) indicates text correctly classified as grooming text, true negative (*TN*) denotes text correctly classified as nongrooming text, false positive (*FP*) represents text incorrectly classified as grooming text, and false negative (*FN*) indicates text incorrectly classified as nongrooming text.

4.2 Experimental Design

The model was trained using the PyTorch deep learning framework with functions from the scikit-learn library. OCR refers to the process of converting text images into machine-readable text. The principle of OCR can be broadly divided into three stages:

1. Image preprocessing;
2. Text recognition;
3. Postprocessing.

The first stage involves refining the image through processes such as noise removal, binarization, and scaling. The second stage entails deciphering the characters in the image, which constitutes the core of OCR. Various algorithms can be used for this process, and deep learning techniques have been employed in this study. Finally, the third stage involves refining the recognized characters and extracting the necessary information.

Based on methods introduced by Maliński et al. [25], experiments were conducted to verify the performance of various OCR engines in correctly recognizing characters in integrated circuit (IC) packages. Open-source OCR engines, such as Tesseract, General OCR (GOCR), CuneiForm, Kraken, and A9T9, were tested for this purpose. The performance of these engines was evaluated in terms of the Levenshtein

distance, which quantifies the difference between two sequences as a string metric. In the case of OCR, the Levenshtein distance (Eq. (5)) measures the minimum number of single-character edits (insertions, deletions, or substitutions) required to convert the output text into the correct string [26]. Table 6 presents the test results as an evaluation of the OCR engine performance [25].

$$lev_{a,b}(i,j) = \begin{cases} \max(i,j) & if \min(i,j) = 0 \\ min \begin{cases} lev_{a,b}(i-1,j)+1 \\ lev_{a,b}(i,j-1)+1 \\ lev_{a,b}(i-1,j-1)+1_{(a_i \neq b_j)} \end{cases} & otherwise \end{cases}. \quad (9)$$

Table 6: Performance test results for OCR engine

OCR engine	Average Levenshtein distance
Tesseract	27.63
General OCR (GOCR)	29.33
CuneiForm	38.24
Kraken	31.12
A9T9	36.40

Accurately recognizing characters from IC packages using OCR engines is difficult, which emphasizes the importance of preprocessing stages to enhance image quality and improve OCR performance. In this study, the Python OpenCV library was employed for image preprocessing, followed by text recognition and extraction using the Tesseract engine, which was chosen considering its better testing performance. Fig. 8 illustrates the structure of the Tesseract OCR.

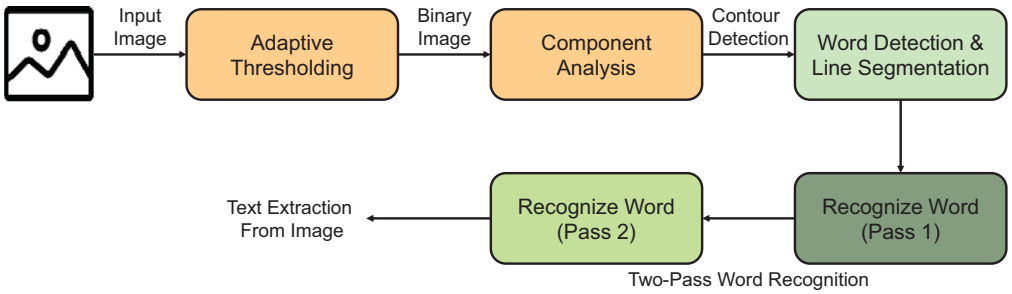


Figure 8: Architecture of Tesseract OCR

4.3 Grooming Text Detection

The dataset was split into training, validation, and testing sets, and the distribution of sentence lengths in each set was examined. Fig. 9 illustrates the sentence-length distribution in the datasets. Based on these results, the maximum length was set to 512, limiting the maximum length of the tokenized sequences. Shorter sequences were padded to ensure uniform length. The divided datasets were tokenized for training and evaluation using the BERTWordPiece tokenizer with a vocabulary size of 30,000 [20]. After setting the hyperparameters, model training and evaluation were conducted. In this experiment, five additional models

were employed alongside KcELECTRA. The model must handle varied sentence lengths. The model was finetuned for a maximum of 10 epochs, with early stops applied to halt training when the loss on the validation set no longer decreased. Cross-entropy loss was employed as the loss function. Other settings included a batch size of 16 for the training and validation datasets, a weight decay of 0.001, and a learning rate of 5×10^{-6}.

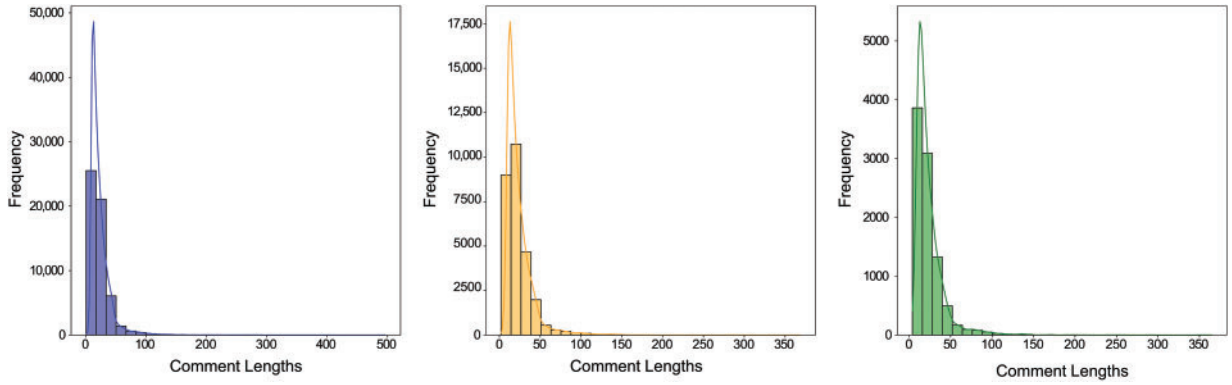


Figure 9: Distribution of sentence lengths in training, validation, and testing datasets. Left (blue): Training set (mostly short sentences <100 tokens); Middle (orange): Validation set (similar pattern, fewer samples); Right (green): Test set (consistent with training and validation sets)

4.3.1 RoBERTa

The RoBERTa [27] model was introduced to address the underfitting problem of the original BERT model. This was achieved by using dynamic instead of static masking to improve performance, which involves changing the masking patterns applied to tokens for each epoch of training data. Additionally, RoBERTa does not perform NSP during training. The experimental results obtained when using complete sentences without NSP loss were better than those obtained when using NSP. Moreover, RoBERTa is pre-trained with larger batch sizes, longer training times, longer sequences, and more data than BERT. The original BERT model uses 30k character-level byte-pair encoding (BPE), whereas RoBERTa uses 50k subword units comprising BPE.

4.3.2 DistilBERT

The DistilBERT [28] model, derived from BERT, applies knowledge distillation techniques. The original BERT model suffers from high memory use and computational overhead; DistilBERT addresses these problems by removing the token-type embeddings and pooler, reducing the number of layers by half. Consequently, the model size is reduced by 40%, and the processing speed is improved by 60% while maintaining 97% of the performance of the original BERT model.

4.3.3 KoBERT

The KoBERT [29] model was developed to overcome the limitations of the existing BERT models in Korean language processing. Comprising the same transformer encoder as BERT, KoBERT is trained on a large-scale corpus consisting of millions of Korean sentences collected from Wikipedia and Korean news sources. By applying data-driven tokenization techniques to reflect the irregular linguistic variations in Korean, KoBERT improves performance by over 2.6%, with only 27% of the tokens in the baseline model.

4.3.4 BLOOMz

The BLOOMz model [30], an improved iteration of the original BLOOM model [31], which stands for BigScience Large Open-science Open-access Multilingual Language Model, was developed through an international collaboration of over a thousand AI experts. The goal was to create a widely accessible large language model (LLM). The original BLOOM model is renowned for its multilingual capabilities and was designed to be freely available, supporting the Open Science initiative to promote transparency and accessibility in AI development. BLOOMz builds on this foundation and is based on the Megatron-LM GPT-2 architecture [32], known for its effectiveness in handling large-scale language models. One of its key advances is the use of ALiBi positional embeddings. Unlike traditional positional embeddings, which rely on absolute positions within sequences, ALiBi embeddings allow for better generalization to sequence lengths not encountered during training, making the model exceptionally adaptable to different text lengths.

BLOOMz is an excellent benchmark model for several reasons. Its ability to support multiple languages and adapt to different text lengths makes it a comprehensive tool for benchmarking complex NLP tasks. Its state-of-the-art architecture sets a high standard for the evaluation of other models. In addition, its innovative use of ALiBi positional embeddings enhances its ability to handle long-range dependencies within text, which is critical for real-world applications involving diverse and large datasets. The model's performance in downstream tasks has shown improvements due to its robust understanding and processing capabilities, establishing it as a rigorous standard for comparison.

Using BLOOMz as a benchmark to evaluate our KcELECTRA model provides us with a clear measure of its effectiveness in handling Korean text, especially in the nuanced environment of Korean SNS conversations. By demonstrating that KcELECTRA can effectively handle specific linguistic challenges and outperform a recognized state-of-the-art model such as BLOOMz, we can solidify the practical utility and technological superiority of our model in the field of NLP tasks. This approach not only reinforces the robustness of KcELECTRA but also underscores our commitment to advancing the capabilities of NLP technologies.

4.3.5 FNet

The FNet model [33] was introduced to enhance the computational efficiency of transformer-based architecture by replacing the traditional self-attention mechanism. Instead of using attention layers, FNet employs the Fourier Transform to capture long-range dependencies within a sequence, significantly reducing computational complexity. This transformation enables efficient global information processing while maintaining strong performance across various NLP tasks. Unlike conventional transformers, FNet removes the quadratic complexity associated with self-attention, leading to faster training and inference speeds while requiring fewer computational resources. Additionally, FNet does not rely on token-to-token attention mechanisms, which can introduce scalability issues in standard transformer models. Furthermore, FNet is pre-trained on large-scale text datasets, ensuring a diverse and representative corpus similar to those used in traditional transformer models. Due to its lightweight architecture, FNet can efficiently handle long sequences while maintaining competitive accuracy, making it particularly well-suited for resource-constrained environments and applications requiring real-time text processing.

4.3.6 DeBERTa-v3

The DeBERTa-v3 model [34], introduced in 2021, was designed as an enhancement over previous transformer-based architectures, addressing key limitations in self-attention mechanisms and computational efficiency. Unlike conventional models such as BERT and RoBERTa, which rely on absolute positional

embeddings, DeBERTa-v3 employs disentangled attention, a mechanism that separately encodes content and positional information to improve representation learning. This enables the model to capture more nuanced contextual relationships within text, leading to improved performance across various NLP tasks. Another major improvement in DeBERTa-v3 is the adoption of ELECTRA-style training, which replaces traditional MLM with a more efficient RTD objective. This approach significantly enhances sample efficiency, allowing the model to achieve superior performance with fewer computational resources. Additionally, DeBERTa-v3 incorporates an optimized pre-training procedure, reducing training costs while maintaining state-of-the-art performance. Furthermore, DeBERTa-v3 is trained with larger batch sizes, extended training durations, and longer sequences compared to its predecessors. While earlier models such as BERT utilized a 30k token vocabulary based on BPE, and RoBERTa expanded this to 50k subword units, DeBERTa-v3 leverages a disentangled embedding approach to refine token representation and improve generalization. These architectural enhancements make DeBERTa-v3-large a robust and efficient model for a wide range of NLP applications, setting a new benchmark for natural language understanding.

4.4 Performance Evaluation

Table 7 compares the performance of the proposed framework and other transformer-based models on the testing set, demonstrating the high accuracy of KcELECTRA under controlled conditions. Meanwhile, Table 8 shows the performance of the model on new datasets, highlighting its adaptability and reliability in different linguistic contexts. The datasets used to evaluate the performance of the KcELECTRA model, as shown in Table 8, were comprehensively selected and prepared to cover a wide range of communication styles and topics prevalent in Korean SNS. This preparation involved several critical steps to ensure the integrity and reproducibility of the research results.

Table 7: Comparison of accuracy with other models

Model	Accuracy	Precision	Recall	F1-score
RoBERTa	0.934	0.930	0.929	0.929
DistilBERT	0.582	0.559	0.536	0.527
KoBERT	0.580	0.539	0.526	0.518
BLOOMz	0.902	0.897	0.894	0.895
FNet	0.891	0.886	0.883	0.884
DeBERTa-v3	0.902	0.895	0.896	0.895
KcELECTRA	0.953	0.948	0.943	0.945

Table 8: Comparative evaluation of model performance using datasets for general chat, sexual remarks, and hate speech in Korean

Model	Accuracy	Precision	Recall	F1-score
RoBERTa	0.796	0.814	0.796	0.798
DistilBERT	0.306	0.352	0.306	0.293
KoBERT	0.283	0.314	0.283	0.279
BLOOMz	0.763	0.772	0.763	0.765
FNet	0.616	0.676	0.616	0.629
DeBERTa-v3	0.736	0.763	0.736	0.740

(Continued)

Table 8 (continued)

Model	Accuracy	Precision	Recall	F1-score
KcELECTRA	0.853	0.862	0.853	0.854

First, a variety of sources were used to collect data that could accurately represent typical and atypical language use in different contexts. General chat logs were obtained from anonymized conversations on KakaoTalk, reflecting a variety of demographic backgrounds. To address the lack of readily available data on sexually explicit or grooming-related communication, the PJZC dataset, known for its examples of potentially harmful online interactions, was translated into Korean with careful attention to linguistic nuances through a combination of machine translation and human review processes. For hate speech, a curated dataset from Kaggle was used, which includes various instances of discriminatory language within the Korean online community.

Each dataset was carefully preprocessed to standardize the format and remove any personally identifiable information, adhering to strict privacy and ethical standards. Expert linguists and domain specialists annotated the data, categorizing each entry into one of the relevant study categories: general chat, sexual remarks, and hate speech. To evaluate the model's performance, 100 sentences from each category were randomly selected to ensure a fair representation of each text type. This stratification helps reduce selection bias and improves the robustness of the model's evaluation. Contrary to typical training, validation, and testing splits, we did not split these datasets. Instead, we tested the model's adaptability and generalization capabilities by applying it to this new, unseen data, considering these 300 samples (100 from each category) as a single test set to assess how well the pre-trained model could handle new and varied inputs without retraining.

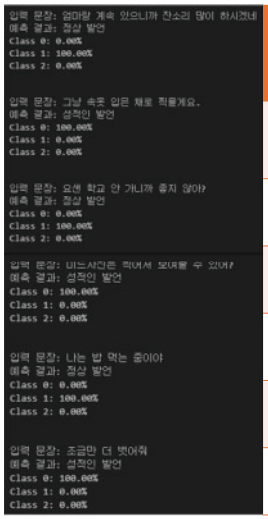
The superior performance of the KcELECTRA model, as shown in Tables 7 and 8, can be attributed to several key factors that enhance its learning efficiency and generalizability across different datasets. This detailed analysis aims to explain these factors and ensure that there is no ambiguity regarding the observed results.

1. Advanced pre-training techniques: The KcELECTRA model employs a discriminative pre-training method where it learns to distinguish between correctly predicted tokens and intentionally corrupted tokens. This approach not only refines its ability to understand the context within Korean SNS conversations but also improves its sensitivity to nuances in different text types.
2. Optimized model architecture: KcELECTRA's architecture has been carefully optimized for the Korean language, incorporating character-level embeddings that capture intricate linguistic features unique to Korean syntax and semantics. This optimization allows for a more accurate interpretation of colloquial expressions and slang common in online grooming conversations.
3. Robust tuning on diverse data: Unlike models trained on narrow datasets, KcELECTRA has been fine-tuned on a wide range of text sources, including general chat logs, sexual comments, and hate speech. This diverse training significantly reduces the risk of overfitting by exposing the model to a wide range of language styles and topics, thereby improving its ability to generalize to unseen data.
4. Regularization and dropout techniques: To further protect against overfitting, sophisticated regularization strategies have been implemented, including dropout layers within the neural network. These techniques help prevent the model from relying too heavily on any single feature or pattern observed in the training data, promoting more robust performance in real-world applications.

5. Continuous evaluation and feedback: Throughout the development phase, the model was continuously evaluated using cross-validation methods across different data splits. This iterative process allowed for constant tuning and refinement of the model based on performance feedback, ensuring that each version of KcELECTRA was an improvement over the previous one.

The collective impact of these strategies is evident in the KcELECTRA model's high accuracy, precision, recall, and F1-score on challenging datasets. This comprehensive approach not only ensures the model's effectiveness in detecting grooming behavior but also its adaptability to evolving language patterns on social platforms. By presenting these technical details, we have provided a clear understanding of why KcELECTRA outperforms other models, supporting its potential for broader implementation in digital security tools without misrepresenting its capabilities.

Fig. 10 shows the results of processing novel text with the constructed model, while Fig. 11 shows the confusion matrix results for the same model, demonstrating its performance across different text classifications. The confusion matrices shown in Fig. 11 represent the performance of five different models in classifying text into three categories: sexually explicit, normal speech, and hate speech. The matrices for RoBERTa, DistilBERT, KoBERT, BLOOMz, FNet, DeBERTa-v3, and KcELECTRA are shown from (a) to (g). RoBERTa has high accuracy across all categories but struggles to distinguish between sexually explicit content and normal speech, as well as between hate speech and normal speech. This suggests a tendency to misclassify moderate expressions within these classes. DistilBERT shows a balanced approach with substantial true positives, especially in normal speech, although it occasionally misclassifies sexually explicit and hate speech as normal.



Sentence (K: Korean Conversation, E: English Translation)	Classification
K: 엄마랑 계속 있으니깐 잔소리를 많이 하시겠네 E: Because I'm still with my mom, I get nagged a lot.	Normal
K: 그냥 속옷 입은 채로 찍을게요. E: I'lljust take a picture with my underwear on.	Sexual
K: 요샌 학교 안 가니깐 좋지 않아? E: Isn'tit nice that you don't go to school these days?	Normal
K: 미드사진은 찍어서 보여줄 수 있어? E: Canyou take a picture of your breasts and show me?	Sexual
K: 나는 밥 먹는 중이야 E: I'meating rice.	Normal
K: 조금만 더 벗어줘 E: Takeoff a little more	Sexual

Figure 10: Classification results for input text

KoBERT excels especially in normal speech, with fewer correct predictions in the sexually explicit and hate speech categories and some overlap between sexually explicit and normal speech. BLOOMz achieves strong performance in recognizing normal speech and effectively classifies both sexually explicit and hate speech with minimal misclassification, indicating a more refined separation between classes. FNet, as shown in its confusion matrix, demonstrates competitive performance, particularly in normal speech. However, it exhibits a slightly higher misclassification rate for sexually explicit content and hate speech compared to some

of the other models. This indicates that while FNet effectively handles structured language, it may struggle slightly with nuanced or context-dependent classifications.

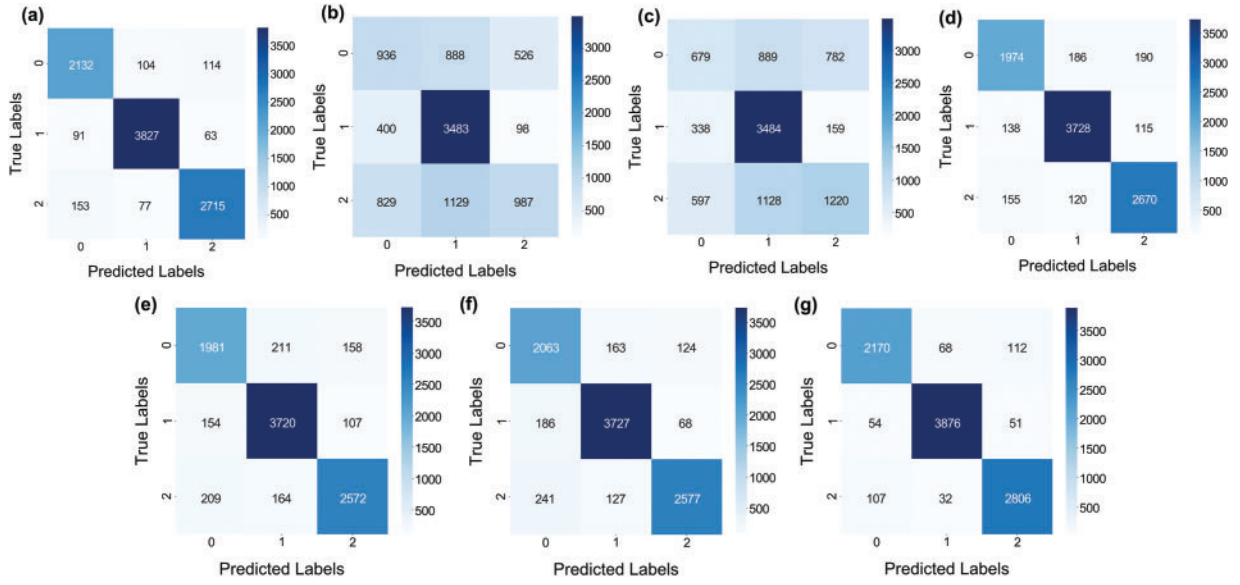


Figure 11: Confusion matrix results for various models: RoBERTa in (a), DistilBERT in (b), KoBERT in (c), BLOOMz in (d), FNet in (e), DeBERTa-v3 in (f), and KcELECTRA in (g), with categories labeled as 0 for sexually explicit, 1 for normal speech, and 2 for hate speech. Darker shades indicate a higher number of correct classifications, while lighter shades represent misclassified instances. The confusion matrices highlight the strengths and weaknesses of each model in distinguishing between the three categories, providing insights into misclassification patterns and areas for potential improvement

DeBERTa-v3 shows strong classification accuracy across all categories, with particularly high performance in normal speech. Compared to FNet, DeBERTa-v3 has fewer misclassifications in sexually explicit content and hate speech, suggesting that it better distinguishes subtle linguistic differences in these categories. KcELECTRA shows excellent performance in all categories with very little misclassification, suggesting a robust ability to discriminate between the three types of speech. The shade of blue in each cell indicates the number of predictions, with darker shades showing higher numbers, visually illustrating the model's effectiveness and areas of potential confusion. This visualization helps you quickly assess the accuracy and reliability of each model for different types of speech.

Analyzing the misclassification patterns in the confusion matrix reveals the challenges associated with handling borderline cases. For example, the KcELECTRA model developed in this study classified the sentence "*Women should always be obedient*" as sexually explicit content, even though it could also be interpreted as hate speech targeting a specific gender. This misclassification illustrates the inherent difficulty in distinguishing between sexually suggestive statements and gender-discriminatory remarks, as both may share similar linguistic structures and contextual nuances, making precise classification challenging. Such ambiguity in classification emphasizes the need for a more refined contextual analysis and greater model robustness, as these subtle overlaps can contribute to errors.

T-distributed stochastic neighbor embedding (t-SNE) is a nonlinear dimensionality reduction technique specialized for visualizing high-dimensional data in a lower-dimensional space, typically 2D or 3D [35]. This technique is designed to preserve the distribution of data in the high-dimensional space

while maintaining local similarities in the lower-dimensional representation, making it particularly useful for detecting clusters and relationships within complex datasets [36]. In this study, t-SNE was applied to intuitively analyze the feature space learned by the KcELECTRA model. This visualization allows for a clearer understanding of how the model structures data across different categories and provides a qualitative assessment of its classification performance.

Fig. 12 presents a t-distributed stochastic neighbor embedding (t-SNE) visualization that effectively illustrates how the KcELECTRA model categorizes textual data into three distinct classes: sexually explicit content (Class 0), normal speech (Class 1), and hate speech (Class 2). This dimensionality reduction technique provides an interpretable representation of the model's learned feature space, allowing for a clear assessment of its clustering capabilities. In the visualization, Class 0 (sexually explicit content) is predominantly represented in blue, forming a well-defined cluster primarily positioned on the left side of the plot. This distinct separation suggests that the model exhibits strong discriminatory power in identifying sexually explicit language. Class 1 (normal speech), shown in green, is distributed across the middle and right sections of the plot, with some degree of overlap with Class 0 and Class 2. This partial intersection reflects the linguistic continuity between neutral speech and more extreme forms of expression. Finally, Class 2 (hate speech), depicted in yellow, is primarily concentrated at the top of the visualization, maintaining a relatively distinct boundary from Class 0 while showing a slight overlap with Class 1. The observed clustering patterns highlight the KcELECTRA model's ability to effectively differentiate between different categories of text, despite the inherent complexity and subtle nuances of human language. This visualization serves as a qualitative assessment of the model's classification performance, demonstrating its capacity to capture meaningful distinctions in textual data. Future refinements could explore methods to further enhance boundary separability, particularly in areas where category overlap is observed.

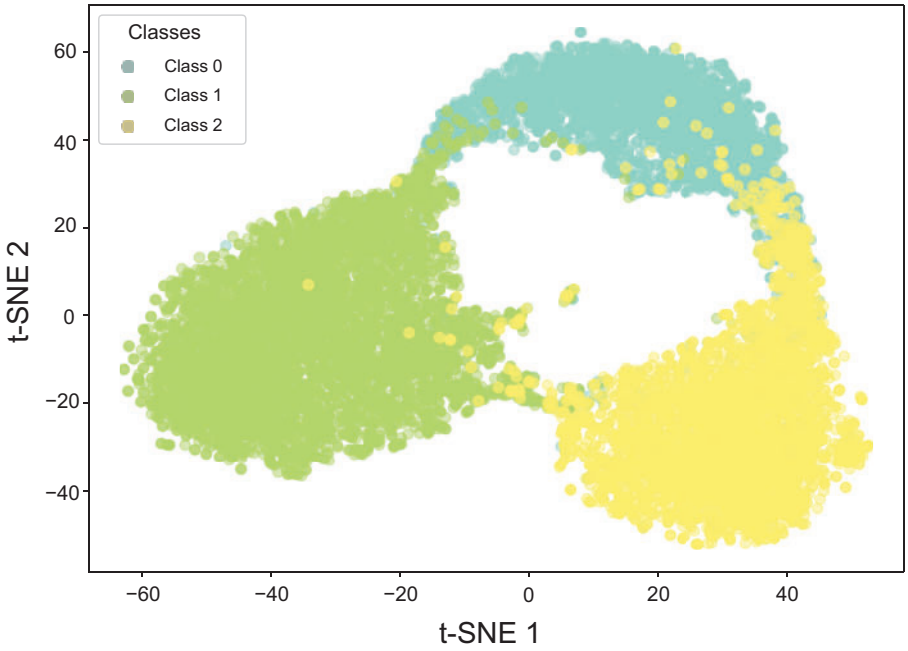


Figure 12: t-SNE visualization of text categorization results using the KcELECTRA model. The three distinct clusters represent different text categories: Class 0 (sexually explicit content, shown in blue), Class 1 (normal speech, shown in green), and Class 2 (hate speech, shown in yellow)

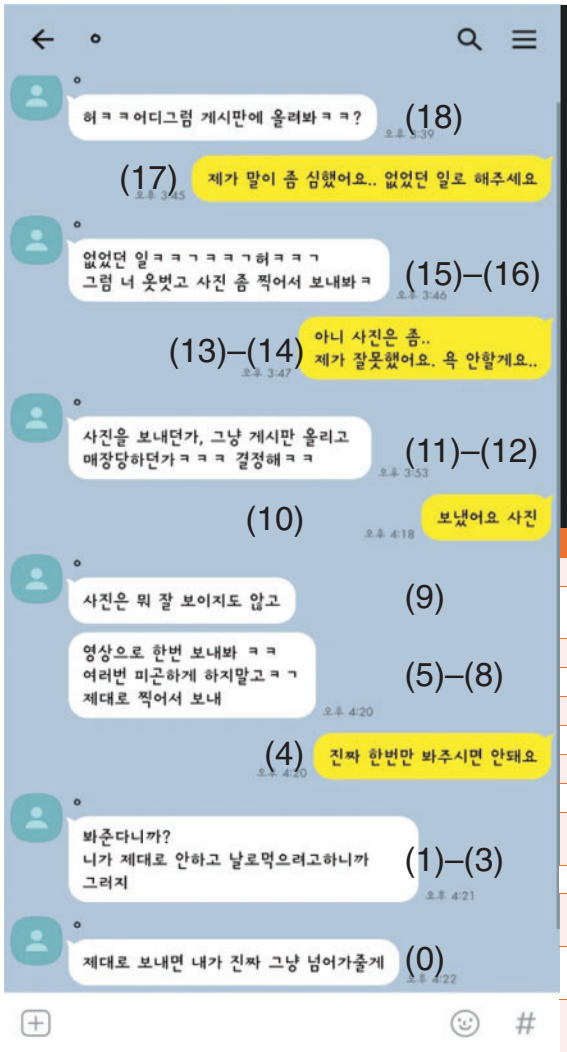
Table 9 presents a comprehensive comparative analysis of the computational requirements and performance metrics of different models. This analysis not only highlights the trade-offs between computational intensity and model accuracy but also demonstrates the strategic advantages of the KcELECTRA model in operational settings. Notably, these evaluations were conducted on a typical consumer-level system, similar to those found in many small technology startups, as mentioned earlier. This ensures that our findings are applicable to environments that do not have dedicated high-performance computing resources.

Table 9: Comparative analysis of model parameters and computational time

Model	Parameters	Time per epoch
RoBERTa	125 million	Approximately 11 min
DistilBERT	66 million	Approximately 6 min
KoBERT	92 million	Approximately 5 min
BLOOMz	560 million	Approximately 14 h
FNet	83 million	Approximately 10 min
DeBERTa-v3 (large)	305 million	Approximately 10 h
KcELECTRA	110 million	Approximately 8 min

The KcELECTRA model stands out not only for its impressive accuracy, achieving a high score of 0.953 in online grooming content detection, but also for its computational efficiency. Compared to more parameter-intensive models such as BLOOMz, which require extensive processing time, KcELECTRA strikes an optimal balance between performance and computational resource usage. This balance ensures that KcELECTRA is both powerful and practical, making it suitable for use in diverse environments, especially where resources may be limited. Our approach carefully balances advanced technological requirements with resource availability, focusing on the operational feasibility of models in different environments. By prioritizing models based on both their performance metrics and their adaptability to different technological capabilities, we enhance the scalability of our grooming detection technology. This strategic selection enables broader deployment and facilitates access even in regions where technological resources are scarce.

After being extracted using OCR algorithms, the text was post-processed before being applied to the constructed model. The postprocessing involved removing time output sections and eliminating parts with only one character. The post-processed text was compared with the text in the original image based on the Levenshtein distance. An average distance of 23.72 indicates improved performance compared with the conventional Tesseract engine. Fig. 13 illustrates the results obtained after applying the post-processed text to the model. The proposed KcELECTRA model achieved a high accuracy of 0.953 in detecting online grooming text, demonstrating efficient learning and excellent predictive performance across data types. The OCR algorithms also exhibited outstanding performance on the extracted text applied to the model, underscoring the potential applicability of OCR in detecting online grooming. Moreover, the integration of OCR algorithms allows the model to rapidly process input images, extracting and classifying text within approximately 10 s, highlighting its practical usability. In addition, the OCR algorithms demonstrated strong performance when applied to the extracted text, further validating their effectiveness in detecting online grooming content.



Sentence Prediction		
0	제대로 보내면 내가 진짜 그냥 넘어가줄게	성적인 발언
1	봐준다니까?	정상 발언
2	니카 제대로 안하고 날로먹으려고하니까	정상 발언
3	그러지	정상 발언
4	진짜 한번만 봐주시면 안돼요	정상 발언
5	영상으로 한번 보내봐 ㅋㅋ	정상 발언
6	영상 한번	정상 발언
7	여러번 피곤하게 하지말고 ㅋㄱ	정상 발언
8	제대로 찍어서 보내	정상 발언
9	사진은 뭐 잘 보이지도 않고	정상 발언
10	보냈어요 사진	정상 발언
11	사진을 보내던카, 그냥 게시판 올리고	정상 발언
12	매장당하던가ㅋㅋㅋ 결정해ㅋㅋ	정상 발언
13	아니 사진은 좀.....	혐오 발언
14	제가 잘못했어요. 욕 안할게요..	정상 발언
15	없었던 일ㅋㅋㄱㅋㅋㅋㄱ혀ㅋㅋㅋ	정상 발언
16	그럼 너 옷벗고 사진 좀 찍어서 보내봐ㅋ	성적인 발언
17	제카 말이 좀 심했어요.. 없었던 일로 해주세요	정상 발언
18	허ㅋㅋ어디그럼 게시판에 올려봐ㅋㅋㅎ	정상 발언

Sentence	Classification
Send it properly, and I'll really let it slide this time.	Sexual
I said I would look at it, didn't I? It's because you're trying to get off easy.	Normal
Can't you just let it go this one time, please?	Normal
The photo doesn't show much. Send a video,	Normal
and don't make me ask again LOL. Do it right.	Normal
The photo doesn't show much. Send a video,	Normal
I couldn't see picture well.	Normal
I sent the photo.	Normal
Send the picture, or just post it yourself and get buried in shame LOL. Make up your mind.	Normal
Not the picture, please...	Hate
Pretend it didn't happen, LOL. Now, strip and send me a pic, come on.	Sexual
I spoke too harshly... Please, can we pretend it didn't happen?	Normal
Haha, go ahead and post it on the forum then, why don't ya? LOL	Normal

Figure 13: Text classification results using OCR

4.5 Discussion

In this section, we discuss how our approach differs from previous studies and highlight how it addresses their limitations. We also explore the ethical challenges that may occur with the use of AI-based tools and describe how we mitigate these concerns.

Kamar et al. [37] highlighted that the increase in online grooming, especially targeting minors, requires urgent attention and showed that a lack of parental supervision can exacerbate associated risks. However, their study focused solely on experimental chatbot conversations in controlled settings, whereas our study utilizes large-scale, real-world social media data to better reflect the dynamics of online grooming. Similarly, Gámez-Guadix et al. [38] proposed online grooming prediction based on personality profiling, but this approach can be unreliable and inaccurate due to inconsistencies between assessed traits and behaviors and raises ethical concerns regarding privacy, stereotyping, and bias. To overcome these limitations, our NLP-based framework focuses on linguistic cues within conversations, offering a more objective and context-sensitive assessment. Additionally, analyzing only linguistic patterns without inferring personality traits mitigates privacy concerns and reduces biases [39].

While Quayle et al. [40] researched how online environments play a role in the development of criminal behavior and evasion of detection, our study is different in that it focuses on AI-driven early detection of grooming incidents to enhance preventive measures. Advances in NLP models have enabled effective early intervention by analyzing linguistic patterns to quickly identify subtle clues that are difficult for humans to perceive [41]. Kim et al. [42] analyzed the relationship between body image concerns and social grooming on platforms such as Facebook and highlighted how gender-based differences in body image perception can introduce biases in automated detection. To effectively address issues such as online grooming, cyberbullying, and sexting, it is crucial to eliminate gender-based biases in detection models [43]. Therefore, we developed a gender-neutral model that focuses solely on content, improving adaptability across different digital environments while ensuring fairer and more accurate detection.

In addition to addressing the methodological limitations of previous studies, we also recognize the importance of using AI-based detection systems ethically and responsibly when utilizing real-world social media data [44]. Rapid technological deployments often overlook crucial aspects of data privacy and security, increasing the risk of exposing personally identifiable information. To address this, our study adheres to rigorous de-identification protocols, ensuring the confidentiality and anonymity of the data subjects. Moreover, while AI-driven real-time monitoring can significantly enhance the detection and prevention of online grooming, it also raises concerns regarding surveillance and misuse. Unregulated AI-based monitoring systems risk privacy violations, necessitating strict oversight, transparency, and data retention limits. To ensure responsible deployment, our research follows stringent privacy regulations and ethical standards. By strictly analyzing linguistic patterns and contextual markers without extracting user identities or metadata, our approach balances effective monitoring with privacy protection, upholding ethical research practices.

Moreover, by bridging text-based and image-based classification, our framework introduces a new perspective for tackling diverse online threats, such as cyberbullying or harassment.

First, we will enhance multimodal detection by integrating NLP for text analysis and computer vision for image classification. Grooming often extends beyond textual manipulation to coercing victims into sharing explicit images, leading to secondary victimization. Thus, detecting both inappropriate text and images is essential for preventing further exploitation. To achieve this, future research will focus on building and optimizing multimodal detection frameworks that leverage both textual and visual data, ensuring more effective identification of grooming behaviors. Additionally, improving OCR performance is critical, as current systems with an average accuracy of 70% struggle with noisy images, limiting their effectiveness in grooming detection. By refining OCR technology alongside advancements in multimodal detection, future efforts will enhance text recognition accuracy in social media conversations, further strengthening the overall reliability of grooming detection frameworks.

Second, we will address data privacy concerns and establish ethical data collection methods. Due to privacy restrictions, collecting large-scale real-world grooming conversation images is both time-consuming and costly. As a short-term solution, we propose leveraging synthetic data generation using GANs or diffusion models, along with image augmentation techniques, to enhance detection capabilities while preserving user privacy. In the long term, we aim to foster ethical collaborations with social media platforms to enable anonymized data collection that complies with privacy regulations. Moreover, through industry partnerships, we seek not only to acquire relevant data but also to implement and evaluate our proposed technology on SNS platforms, ensuring its practical effectiveness in real-world grooming detection.

Third, we will enhance context-aware classification performance for more accurate and reliable detection. To improve the contextual understanding of grooming conversations, we plan to refine our transformer-based models by incorporating longer conversational sequences, enabling the model to better

capture nuanced speech patterns and contextual dependencies. Despite its promising performance, the current model's reliance on shorter sequences limits its ability to fully capture the complexity and evolution of grooming conversations over extended interactions. This enhancement will improve speaker intent recognition, emotion analysis, and dialogue modeling, leading to a more comprehensive assessment of grooming behaviors. Additionally, we plan to integrate psychological models and communication theories, such as social grooming theory, to gain deeper insights into interpersonal interactions, ultimately enhancing classification accuracy. Furthermore, expanding the application of this technology to e-commerce platforms could contribute to online security by detecting fraudulent reviews and inappropriate content. These advancements would extend the impact of this study beyond SNS safety, addressing broader challenges in digital security.

5 Conclusion

Advances in technology have made online communication more accessible, but they have also heightened the risk of minors being exploited by adults. Detecting grooming-related conversations is essential for protecting children and adolescents. In South Korea, research on online grooming has primarily focused on legal and institutional aspects, with limited studies addressing the direct detection of grooming conversations. This study fills this gap by introducing a novel framework that integrates OCR technology with NLP techniques, enabling effective analysis of text embedded within SNS chat images and surpassing the performance of existing Korean transformer-based models.

However, certain limitations remain. Our OCR accuracy, averaging around 70%, can be insufficient for reliably capturing subtle or ambiguous expressions typical in grooming-related conversations. In addition, data scarcity—due to privacy concerns and the limited availability of real-world grooming conversation images—may affect the generalizability of our findings. Finally, our confusion matrix revealed a tendency to misclassify sexually explicit speech and hate speech because of their semantic similarities, highlighting the need for more nuanced, context-aware classification.

Despite these constraints, our framework demonstrates competitive efficiency and accuracy in detecting grooming incidents, thus contributing to safer online environments. Focusing solely on linguistic cues and de-identified data, our model emphasizes ethical considerations and strives to balance effective monitoring with respect for user privacy. Overall, this study underscores the importance of ongoing advancements in AI-based solutions to guard against online grooming and other digital threats. Future efforts will aim to refine multimodal detection, address privacy concerns, and enhance context-aware classification to further improve the robustness and ethical deployment of grooming detection systems.

Acknowledgement: Not applicable.

Funding Statement: This work was supported by the IITP (Institute of Information & Communications Technology Planning & Evaluation)–ITRC (Information Technology Research Center) grant funded by the Korean government (Ministry of Science and ICT) (IITP-2025-RS-2024-00438056).

Author Contributions: Conceptualization, Sangmin Kim and Byeongcheon Lee; methodology, Sangmin Kim; software, Sangmin Kim and Byeongcheon Lee; validation, Muazzam Maqsood and Jihoon Moon; formal analysis, Muazzam Maqsood; investigation, Byeongcheon Lee and Muazzam Maqsood; resources, Sangmin Kim and Byeongcheon Lee; data curation, Sangmin Kim; writing—original draft preparation, Sangmin Kim; writing—review and editing, Jihoon Moon and Seungmin Rho; visualization, Sangmin Kim and Jihoon Moon; supervision, Jihoon Moon and Seungmin Rho; project administration, Seungmin Rho; funding acquisition, Seungmin Rho. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: This study uses Korean-specific datasets from AI-Hub, including Korean SNS data, text ethics verification, and hate speech, all formatted in JSON. These datasets cover everyday conversations and unethical sentiments with anonymized personal information. These resources are designed for research on Korean social media and ethics. Available at https://www.aihub.or.kr/aihubdata/data/view.do?dataSetSn=114 and https://www.aihub.or.kr/aihubdata/data/view.do?dataSetSn=558 (accessed on 1 April 2025).

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Hong JY, Shim SJ, Lee SJ. Random chat application conversation content analysis using topic modeling (LDA): focusing on cases of online grooming of adolescents. Korean Criminal. 2023;17(1):5–26. doi:10.29095/JKCA.17.1.1.
2. Welner M. Child Sexual Abuse: 6 Stages of Grooming; 2020. [cited 2023 Aug 10]. Available from: https://www.oprah.com/oprahshow/child-sexual-abuse-6-stages-of-grooming/all.
3. Oh SY, Shin HJ. A study on online grooming sex offenses status and victim protection. Police Sci J. 2019;14(4):133–60.
4. Koreabizwire. Open chats leave teenagers at risk of online grooming; 2024. [cited 2024 Mar 20]. Available from: http://koreabizwire.com/open-chats-leave-teenagers-at-risk-of-online-grooming/210623.
5. Wikipedia. Nth Room case; 2024. [cited 2024 Mar 20]. Available from: https://en.wikipedia.org/wiki/Nth_Room_case.
6. Yonhap News Agency. Police officers to go undercover to chase online sexual predators; 2021. [cited 2024 Mar 20]. Available from: https://en.yna.co.kr/view/AEN20210923008600315.
7. Kheddar H. Transformers and large language models for efficient intrusion detection systems: a comprehensive survey. arXiv:2408.07583. 2024.
8. Milon-Flores DF, Cordeiro RL. How to take advantage of behavioral features for the early detection of grooming in online conversations. Knowl Based Syst. 2022;240(C):108017. doi:10.1016/j.knosys.2021.108017.
9. Jeon S, Kwon H, Keong H, Kim S. Online sexual grooming offenses against children and adolescents. Korean J Soc Welf Stud. 2021;52(1):97–138.
10. Anderson P, Zuo Z, Yang L, Qu Y. An intelligent online grooming detection system using AI technologies. In: Proceedings of the IEEE International Conference on Fuzzy Systems (FUZZ-IEEE); 2019; New Orleans, LA, USA. p. 1–6. doi:10.1109/FUZZ-IEEE.2019.8858973.
11. Bours P, Kulsrud H. Detection of cyber grooming in the online conversation. In: Proceedings of the IEEE International Workshop on Information Forensics and Security (WIFS); 2019; Rome, Italy. p. 1–6. doi:10.1109/WIFS47025.2019.9035090.
12. Guo Z, Wang P, Huang L, Cho JH. Authentic dialogue generation to improve youth's awareness of cyber grooming for online safety. In: Proceedings of the IEEE 35th International Conference on Tools with Artificial Intelligence (ICTAI); 2023; Atlanta, GA, USA. p. 64–9. doi:10.1109/ICTAI59109.2023.00017.
13. NIA AI-Hub. Korean SNS. 2024. [cited 2024 Mar 20]. Available from: https://www.aihub.or.kr/aihubdata/data/view.do?dataSetSn=114.
14. NIA AI-Hub. Text ethics verification data. 2024. [cited 2024 Mar 20]. Available from: https://www.aihub.or.kr/aihubdata/data/view.do?dataSetSn=558.
15. Lee J, Lim T, Lee H, Jo B, Kim Y, Yoon H, et al. K-MHaS: a multi-label hate speech detection dataset in korean online news comment. In: Proceedings of the 29th International Conference on Computational Linguistics; 2022 Oct 12–17; Gyeongju, Republic of Korea. p. 3530–8.
16. Ludobic. KakaoChatData. 2024. [cited 2024 Aug 13]. Available from: https://github.com/Ludobico/KakaoChatData.
17. Danielafe7-usp. PJZC dataset. 2024. [cited 2024 Aug 13]. Available from: https://github.com/danielafe7-usp/BF-PSR-Framework/tree/main/JsonData.

18. Tanat05_2024. Korean-hate-chat-data. 2024. [cited 2024 Aug 13]. Available from: https://www.kaggle.com/datasets/tanat05/korean-hate-chat-data.
19. Yıldız K, Yıldız Z, Demir Ö., Buldu A. Determination of yarn twist using image processing techniques. In: International Conference on Image Processing, Production and Computer Science; 2015; Istanbul, Turkey. p. 83–8.
20. KcELECTRA. Korean comments ELECTRA. 2022. [cited 2024 Mar 10]. Available from: https://github.com/Beomi/KcELECTRA.
21. Park L. NSMC. 2023. [cited 2024 Mar 20]. Available from: https://github.com/e9t/nsmc.
22. Naver Corporation. NLP Challenge-Named Entity Recognition (NER). 2024. [cited 2024 Mar 20]. Available from: https://github.com/naver/nlp-challenge/tree/master/missions/ner.
23. Zhang Y, Baldridge J, He L. PAWS: paraphrase adversaries from word scrambling. In: Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL); 2019 Jun 2–7; Minneapolis, MN, USA.
24. Ham J, Choe YJ, Park K, Choi I, Soh H, et al. KorNLI and KorSTS: new benchmark datasets for Korean natural language understanding. arXiv:2004.03289. 2020. [cited 2024 Mar 10]. Available from: https://arxiv.org/abs/2004.03289.
25. Maliński K, Okarma K. Analysis of image preprocessing and binarization methods for OCR-based detection and classification of electronic integrated circuit labeling. Electronics. 2023;12(11):2449. doi:10.3390/electronics12112449.
26. Yujian L, Bo L. A normalized Levenshtein distance metric. IEEE Trans Pattern Anal Mach Intell. 2007;29(6):1091–5. doi:10.1109/TPAMI.2007.1078.
27. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: a robustly optimized BERT pretraining approach. arXiv:1907.11692. 2019.
28. Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv:1910.01108. 2019.
29. SKTBrain. KoBERT. 2024. [cited 2024 Mar 20]. Available from: https://github.com/SKTBrain/KoBERT.
30. Workshop B, Scao TL, Fan A, Akiki C, Pavlick E, Ilić S, et al. Bloom: a 176b-parameter open-access multilingual language model. arXiv:2211.05100. 2022.
31. Muennighoff N, Wang T, Sutawika L, Roberts A, Biderman S, Scao TL, et al. Crosslingual generalization through multitask finetuning. arXiv:2211.01786. 2022.
32. Shoeybi M, Patwary M, Puri R, LeGresley P, Casper J, Catanzaro B. Megatron-LM: training multi-billion parameter language models using model parallelism. arXiv:1909.08053. 2019.
33. Lee-Thorp J, Ainslie J, Eckstein I, Ontanon S. Fnet: mixing tokens with fourier transforms. arXiv:2105.03824. 2021.
34. He P, Gao J, Chen W. DeBERTaV3: improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. arXiv:2111.09543. 2021.
35. Van der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res. 2008;9(11):2579–605.
36. Yildiz K, Çamurcu AY, Dogan B. Comparison of dimension reduction techniques on high dimensional datasets. Int Arab J Inf Technol. 2018;15(2):256–62.
37. Kamar E, Maimon D, Weisburd D, Shabat D. Parental guardianship and online sexual grooming of teenagers: a honeypot experiment. Comput Human Behav. 2022;137(11):107386. doi:10.1016/j.chb.2022.107386.
38. Gámez-Guadix M, Mateos-Pérez E. Longitudinal and reciprocal relationships between sexting, online sexual solicitations, and cyberbullying among minors. Comput Human Behav. 2019;94(1):70–6. doi:10.1016/j.chb.2019.01.004.
39. Choudhary A, Arora A. Linguistic feature based learning model for fake news detection and classification. Expert Syst Appl. 2021;169(2):114171. doi:10.1016/j.eswa.2020.114171.
40. Quayle E, Allegro S, Hutton L, Sheath M, Lööf L. Rapid skill acquisition and online sexual grooming of children. Comput Human Behav. 2024;39(1–2):368–75. doi:10.1016/j.chb.2014.07.005.
41. Ringenberg TR, Seigfried-Spellar K, Rayz J. Assessing differences in grooming stages and strategies in decoy, victim, and law enforcement conversations. Comput Human Behav. 2024;152(9):108071. doi:10.1016/j.chb.2023.108071.

42. Kim JW, Chock TM. Body image 2.0: associations between social grooming on Facebook and body image concerns. Comput Human Behav. 2015;48(1):331–9. doi:10.1016/j.chb.2015.01.009.
43. Yokotani K, Takano M. Predicting cyber offenders and victims and their offense and damage time from routine chat times and online social network activities. Comput Human Behav. 2022;128(1):107099. doi:10.1016/j.chb.2021.107099.
44. Hernández MP, Schoeps K, Maganto C, Montoya-Castilla I. The risk of sexual-erotic online behavior in adolescents—which personality factors predict sexting and grooming victimization? Comput Human Behav. 2021;114(2):106569. doi:10.1016/j.chb.2020.106569.