Chapter 14
Enhancing Digital Safety: Exploring Parental Control Mechanisms for Secure Online Experiences

Mukesh Kumar
Advanced Centre of Research & Innovation (ACRI), Department of Computer Application, Chandigarh School of Business, Chandigarh Group of Colleges Jhanjeri, Mohali, 140307, Punjab, India

Mirza Shuja
https://orcid.org/0000-0001-5058-003X
School of Computer Application, Lovely Professional University, Phagwara, Punjab, India

Bilal Ahmed
School of Computer Application, Lovely Professional University, Phagwara, Punjab, India

Tanya Gera
Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India

Deepak Thakur
Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India

Vivek Bhardwaj
https://orcid.org/0000-0002-2288-6987
School of Computer Science and Engineering, Manipal University Jaipur, India

ABSTRACT

As children engage more with digital platforms, they face increasing risks such as cyberbullying, exposure to inappropriate content, and online grooming. Cases like Meta's 2023 investigation into the exploitation of minors highlight the pressing need for stronger online safety tools. This study proposes an AI-based parental control framework that goes beyond basic keyword filtering. Leveraging Natural Language Processing (NLP) and Transformer-based models like BERT and GPT-

DOI: 10.4018/979-8-3373-2716-7.ch014

4, the system enables real-time detection of harmful content and adaptive moderation. Trained using the 2023 Cyberbullying Detection Dataset and YouTube's Open Content Moderation Dataset, the model effectively classifies and responds to harmful interactions on social media and messaging platforms. Testing showed an 89% accuracy rate in identifying harmful content and a 40% boost in real-time intervention over traditional tools. The study underscores the promise of AI-driven safety solutions while stressing ethical use and balanced regulation to protect both privacy and well-being.

1. INTRODUCTION

The rapid growth of internet access among children (Roth et al., 2025) has transformed the way they learn, communicate, and entertain themselves. However, this digital immersion comes with escalating risks, including cyberbullying (Rawat et al., 2025), online grooming, exposure to explicit content, and exploitation. As of 2024, over 62% of children aged 8–16 access the internet daily without supervision, with 38% encountering inappropriate or harmful content at least once per month (Statista, 2024). Alarming events such as Meta's 2023 investigation (Humphry et al., 2025) revealed that its platforms failed to adequately restrict harmful content targeting minors in over 70% of flagged instances, prompting global calls for advanced digital safety (Bhardwaj et al., 2024) (Nahar et al.,2023) (Dhawan et al.,2025) frameworks.

Traditional parental control systems—centered on keyword filtering, manual content blocking (Qu et al., 2025), and screen time management—are increasingly ineffective against today's evolving online threats. These systems lack adaptability, contextual understanding, and real-time responsiveness, limiting their ability to address dynamic digital interactions. In response, researchers and developers are turning to Artificial Intelligence (AI) (Rajavat et al., 2024) and Natural Language Processing (NLP) (Mascari et al.,2025) (Pithawa et al., 2023) to create more intelligent, context-aware parental control tools.Modern AI-driven frameworks leverage transformer-based models, such as BERT (Bidirectional Encoder Representations from Transformers) and GPT-4 (Rawat and Rajavat,2024a), which are capable of processing vast amounts of text data, understanding context, and detecting potentially harmful (Mishra et al., 2024) interactions in real time. These tools are complemented by machine learning classifiers (Rawat and Rajavat,2024b), sentiment analysis engines, and deep neural networks for behavior prediction and content moderation. This study employs datasets including the 2023 Cyberbullying Detection Dataset (Philipo et al., 2024), containing over 50,000 labeled instances of abusive language and harassment, and YouTube's Open Content Moderation Dataset (2023) (AlDahoul et al.,2024), with over 1.2 million user comments annotated for harmful content