Dimension-Expanding MLP in Transformer: Inappropriate Sentences and Paragraph Digital Content Filtering

Ariq Cahya Wardana[1,*], Andi Prademon Yunus[2], Rifki Adhitama[3], Muhammad Abdul Latief [4], Martryatus Sofia[5]

[1,2,3,4,5]*Telkom University, DI Panjaitan No 128 Purwokerto, Banyumas 53145, Indonesia*

(Received: November 15, 2024; Revised: December 21, 2024; Accepted: January 20, 2025; Available online: March 15, 2025)

Abstract

The creation of digital content is a crucial aspect of today's digital environment, enabling individuals and organizations to engage audiences effectively. As digital platforms expand, ensuring the security, professionalism, and appropriateness of user-generated content is essential. This paper introduces a novel content filtering framework that leverages dimension-expanding multilayer perceptrons (MLPs) within Transformer architectures to address these challenges. The proposed framework enhances the model's ability to process high-dimensional features by incorporating intermediate feature transformations that refine contextual representations. By expanding dimensions, the model preserves critical information while reducing noise, leading to improved filtering performance over conventional architectures. Specifically, dimension-expanding MLP layers increase the network's representational capacity by projecting lower-dimensional inputs into higher-dimensional spaces before filtering operations. This allows the model to capture intricate feature relationships and detect subtle patterns in digital content. Additionally, non-linear activation functions and normalization techniques improve generalization across diverse user-generated content, ensuring a more robust and adaptive filtering process. We utilize the inappropriate word corpus dataset of KBM.id, consisting of 10,000 entries classified under the labels normal, less sensitive, and sensitive. Experimental results demonstrate that the proposed model outperforms LSTM and GRU, achieving an accuracy of 0.744 compared to LSTM's 0.712 and GRU's 0.739 under optimal conditions. Moreover, it exhibits greater computational efficiency, requiring only 129.1 GFLOPs compared to LSTM's 1.44 TFLOPs and GRU's 1.08 TFLOPs. These findings highlight the model's ability to balance accuracy and efficiency, making it well-suited for large-scale applications. By integrating advanced Transformer-based architectures, this study enables nuanced contextual understanding, crucial for filtering inappropriate or harmful content. Practical applications include social media content moderation, legal document compliance monitoring, and filtering harmful material in e-learning and gaming platforms. This research advances automated, ethical, and scalable digital content curation, offering an accurate and computationally efficient solution for modern digital platforms.

Keywords: Digital Content Filtering, Transformer, Dimension-Expanding MLP, Contextual Understanding, Inappropriate Sentences and Paragraphs, Model Comparison, Content Moderation Applications

1. Introduction

Digital content encompasses a wide range of information types, including text, audio, video, and graphics, which can be distributed electronically [1]. The evolution of digital platforms has significantly transformed content creation, distribution, and consumption, leading to new challenges in the coordination between content generators [2], [3]. Digital content creation has become a central aspect of the contemporary digital landscape, driven by the need for individuals and organizations to engage audiences effectively. The distribution of digital content on various platforms presents significant advantages and challenges. On the positive side, digital platforms facilitate rapid information dissemination, improve accessibility, and allow individuals and businesses to reach wider audiences without geographical restrictions, encouraging creativity and innovation in content presentation [4]. This democratization of content creation can empower users and improve engagement, particularly in educational and mental health contexts [5]. For example, digital mental health interventions have shown promise in improving psychological well-being among college students, highlighting the potential for positive outcomes through effective content distribution [6]. However, this democratization also introduces challenges related to the proliferation of user-generated content (UGC),

***Corresponding author: Ariq Cahya Wardana (1ariqcahya@telkomuniversity.ac.id)**

DOI: https://doi.org/10.47738/jads.v6i2.627

This is an open access article under the CC-BY license (https://creativecommons.org/licenses/by/4.0/).
© Authors retain all copyrights

which often lacks curation or control, potentially leading to harmful or inappropriate material being disseminated. In contrast, the uncontrolled nature of digital content distribution can lead to serious drawbacks, such as the proliferation of misinformation, copyright infringements, and exposure to harmful content. Excessive engagement with negative content, including cyberbullying and unrealistic beauty standards, can adversely affect mental health by increasing stress and lowering self-esteem [7]. Furthermore, the fragmentation of audiences on various platforms complicates the maintenance of message consistency and effective engagement strategies [8].

The curation of digital content is essential to ensure clarity, accuracy, and adherence to ethical standards before public dissemination. This process minimizes the risks of misinformation and harmful impacts on audiences, which is particularly crucial in today's complex digital landscape [9]. By evaluating content structure, coherence, and tone, creators can deliver messages that are not only engaging, but also responsible and beneficial to their target audience [10]. The role of digital curators extends beyond mere content management; they are tasked with developing strategies that improve audience engagement through interactive experiences and the effective use of social networks [10]. This is vital in an environment where attention is a scarce resource and the competition for user engagement is fierce [11]. Furthermore, machine learning-based curation systems are increasingly utilized to organize and present content, further emphasizing the importance of thoughtful curation in maintaining credibility and trust [12]. Effective digital content curation is a multifaceted process that contributes significantly to the quality and impact of online communications.

Machine learning (ML) and natural language processing (NLP) are essential for automating digital content analysis, providing tools to evaluate coherence, tone, and grammar while identifying offensive or inappropriate material. These technologies analyze vast amounts of data to identify patterns, ensuring that the content is relevant and of high quality [12]. The integration of these techniques in the proposed dimension-expanding MLP Transformer is specifically designed to leverage NLP's ability to extract semantic meaning and ML's capability to optimize learning for nuanced classification tasks. Natural language processing (NLP) powered neural networks can assess various aspects of content, including coherence, tone, and grammar, effectively flagging issues such as misinformation or offensive language [13]. This capability is crucial in maintaining ethical standards in content dissemination. ML facilitates the classification of content by audience, topic, or tone and provides actionable suggestions for improvement, such as simplifying text or improving clarity [14]. By integrating these advanced technologies, content curation becomes not only more efficient but also more responsible, reducing the manual effort required while upholding high standards in the rapidly evolving digital landscape [15].

The multilayer perceptron (MLP) is a type of artificial neural network that is designed for supervised learning tasks such as classification and regression. It consists of multiple layers of neurons, with weighted connections between neurons in consecutive layers [16]. Using the Back-Propagation (BP) algorithm, MLP adjusts the connection weights layer by layer, enabling it to classify non-linearly separable patterns and approximate functions [17]. BP works by computing the gradient of the loss function with respect to each weight by propagating errors back through the network, allowing MLP to learn from the discrepancies between the predicted and actual outputs [18]. This combination of multi-layer architecture and iterative optimization makes MLP a versatile and powerful tool for a wide range of applications in artificial intelligence and machine learning. In content filtering, MLP's architecture is particularly well-suited due to its ability to model complex relationships between input features, enabling it to differentiate nuanced patterns in digital content such as detecting harmful language even within subtle contextual variations.Building on these foundation principles, the Transformers model advances neural network architectures by leveraging self-attention mechanisms to process sequences, eliminating recurrent and convolutional layers while capturing dependencies effectively [19]. Within this architecture, MLPs function as position-wise feedforward networks (FFN), refining the output of self-attention layers through two fully connected layers separated by a non-linear activation function, such as GELU or ReLU. This integration of self-attention and MLPs balances global context awareness with localized feature refinement, enabling Transformers to achieve state-of-the-art performance in various tasks [18].

In the proposed approach, these components are specifically adapted to address the challenges of filtering inappropriate digital content. The self-attention mechanism allows the model to capture both local and global dependencies in the text, identifying contextual cues that might indicate harmful or offensive language. For example, it can recognize when certain words or phrases become inappropriate based on the surrounding context, a capability that is essential for nuanced filtering tasks. The dimension-expanding MLP, integrated as a position-wise feedforward network, further enhances this process by transforming the attention outputs into a higher-dimensional space. This transformation allows for more effective separation of similar yet distinct features, enabling the model to make precise decisions in complex filtering scenarios, such as distinguishing colloquial expressions from harmful language. By combining the global contextual awareness of self-attention with the refined feature representation provided by the MLP, the Transformer becomes highly effective at classifying content with high precision and recall. This approach ensures that inappropriate material is accurately flagged while minimizing false positives, making it ideal for applications like social media moderation, compliance monitoring, and e-learning content curation.

By offering a strong framework for removing offensive lines and paragraphs from digital content, this study seeks to overcome the drawbacks of current curation processes by building on the advancements of dimension-expanding MLPs and their incorporation into Transformer systems. In summary We propose a dimension-expanding MLP within the Transformer architecture for sentence-based digital content filtering. We conduct a comprehensive comparison between the Transformer with dimension-expanding MLP, the standard Transformer, and RNN-based models, including Vanilla RNN, LSTM, and GRU. Additionally, we develop a baseline framework for digital content filtering in Bahasa Indonesia to detect inappropriate content. We also demonstrate the applicability of this framework in digital content curation tasks, such as content categorization, metadata generation, and relevance ranking. By achieving these objectives, this research aims to contribute to the development of automated, ethical, and effective digital content curation.

2. Literature Review

2.1. Digital Content

Digital content encompasses information that is generated, stored, and shared in a digital format. This includes a variety of media types, such as images, audio, video, and interactive components, all accessible via digital devices. “Born digital” content, specifically designed for digital platforms, differs from “turned digital” content that is adapted from traditional print or analog formats [20]. Recognizing this distinction is crucial, as it underscores the changing nature of content creation in the digital era. Examples of digital content include everything from social media posts and blogs to e-books and online courses, illustrating the various methods of dissemination and consumption of information [21]. Research indicates that individuals, including students, show varying degrees of proficiency in digital content creation, affecting their ability to interact effectively with digital media [22]. Consequently, there has been an increase in educational approaches that improve learners 'skills in creating digital content, preparing them to navigate and contribute to the digital environment [23]. The emergence of user-generated content (UGC) has reshaped digital content landscapes, enabling individuals to produce and share their own media. This evolution holds considerable implications for content perception and use, as it democratizes content creation, promoting community participation [24].

2.2. Cyberbullying in Digital Platform

Cyberbullying in today’s digital platform involves using online tools like social networking sites and messaging apps to deliberately harass or harm individuals, often marked by continuous aggression and power imbalances [25]. This includes actions such as sending threatening messages, doxxing, and spreading falsehoods, with anonymity and the lasting nature of online content intensifying its effects. Cyberbullying can cause severe emotional and mental health problems, such as anxiety, depression, and suicidal ideation, studies indicating that about half of these acts are committed by recognizable individuals, often peers [26]. The incidence of cyberbullying has increased with the expansion of digital device usage, particularly among young people, with increased risks observed among frequent users, particularly girls [27]. Implementing preventive strategies, such as nurturing empathy, teaching digital literacy, and establishing safe reporting systems, is critical to tackling this problem. Digital content curation, the methodical process of collecting, organizing, and sharing existing materials, aids in combating information overload by filtering pertinent sources [28], organizing information with context [29], synthesizing actionable insights [28], and effectively presenting it to improve comprehension and practical application.

2.3. Curation in Digital Content

Digital content curation includes the processes of selectively collecting, organizing, commenting on, and distributing previously created digital materials to fulfill certain objectives or cater to specific audiences, providing a unified alternative to creating new content in today's information-saturated environment. To identify high-quality content, use tools such as Feedly or Pocket to separate trustworthy sources from unreliable ones [28]. Next, in the Organization and Contextualization phase, curators arrange content and add notes to improve understanding, especially in educational settings [29]. The sense-making stage involves deriving patterns and insights from various sources of information, promoting critical analysis, and idea connectivity [28]. Finally, in the Sharing and Presentation phase, the curated content is shared in user-friendly formats such as blogs or newsletters, focusing on clarity and ease of use [30]. This strategy is essential to effectively navigate the rapid spread of digital content.

2.4. MLP in Transformer and the Multi-Layer Perceptron Method

Multi-Layer perceptrons (MLPs) play an integral role in the encoder-decoder frameworks of transformer models, which are crucial for computer vision and natural language processing applications. After the attention mechanism establishes the relationships among the input tokens, the MLPs refine the output by applying non-linear transformations, often employing activation functions such as GELU [19]. To maintain a balance between efficiency and representational capacity, contemporary Transformers, including BERT and Vision Transformers (ViTs), adjust MLP architectures for specific tasks such as image classification and semantic embedding (Oxford Academic). In general, MLPs remain versatile and indispensable in deep learning, illustrating their importance in tackling complex problems across diverse language and vision tasks.

3. Methodology

This section describes our proposed method in detail from the data acquisition and preprocessing, the dimension-expanding Multi-Layer Perceptron in Transformer, the experiment setup, and evaluation metrics used to validate the method's performance on inferences. The overview of the system is illustrated in figure 1.

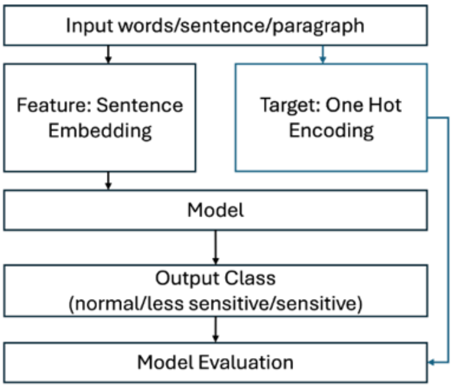

Input words/sentence/paragraph
Feature: Sentence Embedding
Target: One Hot Encoding
Model
Output Class (normal/less sensitive/sensitive)
Model Evaluation

Figure 1. System overview.

3.1. Data Acquisition and Preprocessing

Our dataset annotation utilizes the inappropriate word corpus of KBM.id, consisting of 10,000 entries classified under the labels normal, less sensitive, and sensitive. This corpus was chosen for its relevance to real-world scenarios of content moderation, as it captures diverse linguistic patterns and cultural nuances, making it highly representative of the target content. By using this corpus, the model is trained to effectively identify and filter inappropriate material in a variety of contexts. In addition, we manually annotate the data to accommodate the fairness and human subjectivity. The manual annotation is done by judging the sentence appropriateness based on the expert judgement and is separated from the word-based annotation. The dataset labeling was conducted collaboratively by a team of seven annotators with diverse backgrounds. Each annotator independently reviewed and labeled the data, and any disagreements were resolved through a consensus-based approach. This method ensures that the annotations reflect a balanced perspective, reducing individual biases and enhancing the reliability of the labeled data. To numerically represent the text and associated labels, we use the embedding for the feature and the one-hot encoding technique for the target.

3.1.1. Sentence Embedding

In this research, we applied supervised contrastive learning to enhance sentence embeddings, improving their ability to capture semantic similarities. The approach fine-tunes a pre-trained BERT model by pulling semantically similar sentences closer while pushing dissimilar ones apart [32]. We utilized a pre-trained transformer model from HuggingFace, optimizing it with contrastive loss to achieve better performance on downstream natural language processing tasks.

3.1.2. Positional Encoding

The first step in a Transformer-based model is to apply positional encoding to the input. Since the self-attention mechanism in Transformer models is inherently order-agnostic, positional encoding plays a crucial role in preserving the sequence structure of the text. It works by adding a set of sinusoidal functions to the input embeddings, allowing the model to differentiate between token positions within a sequence. In Transformer-based architectures like BERT and GPT, positional encoding enhances the model's ability to capture sequential relationships between words and phrases, which is essential for various NLP tasks, including sentiment analysis and text classification. The effectiveness of positional encoding lies in its ability to help the model understand dependencies between words that are far apart in the sequence, improving contextual comprehension. When combined with the self-attention mechanism and the dimension-expanding feed-forward layers, positional encoding strengthens the model’s ability to analyze the interplay between word order and meaning. This integration ensures that the model considers not only the content of individual words but also their contextual dependencies, leading to more accurate and meaningful text representations. Given an input sequence, the positional representation of words is determined by adding trigonometric sine and cosine functions to the embedding values under specific conditions, as shown in Equation 1.

$$\vec{P} = \begin{cases} \sin\left(\frac{x_i}{n}\frac{2k}{d}\right) \text{ if } i = 2k \\ \cos\left(\frac{x_i}{n}\frac{2k}{d}\right) \text{ if } i = 2k+1 \end{cases} \tag{1}$$

where n is a scalar determined by default 10000 in the Transformer Networks [33] and k is indices containing {0, 1, . . . , $\frac{d}{2}$ 1}.

3.2. Dimension-expanding Multi-Layer Perceptron

Multi-Layer Perceptron (MLP) is a key component in supervised learning models, including Vision Transformers (ViTs), where it serves as a classifier following the transformer encoder. Unlike convolutional neural networks (CNNs), ViTs rely on self-attention for feature extraction, with MLP layers refining embeddings for image classification. The study highlights MLP’s role in enhancing feature abstraction and classification performance, demonstrating its effectiveness across various benchmarks [34]. In this study, we apply the multilayer perceptron in transformer networks with expanding dimensions. The input dimension is expanded by a much bigger size of the hidden dimension. The main notion to always give the bigger hidden dimension is to consider the output of the MultiHead Attention into higher dimensionality, thus the linear separator in the Linear layer breaks into bigger spatial dimension. The choice to expand the input dimension is rooted in the need to enhance the model’s ability to represent complex patterns and subtle nuances in the data.

By transforming the outputs of the self-attention mechanism into a higher-dimensional space, the model gains additional capacity to disentangle overlapping feature representations, enabling finer-grained distinctions between appropriate and inappropriate content. This dimensional expansion is particularly effective in content filtering because it allows the MLP to capture and emphasize subtle contextual cues, such as the tone or implied meaning of phrases that may not be immediately apparent in lower-dimensional spaces. For instance, the model can better differentiate between offensive and neutral language when the features are mapped into a space that amplifies their differences. Moreover, the increased dimensionality improves the separability of feature representations in the classification process, reducing the likelihood of misclassifications. While this approach increases computational complexity, the benefits in accuracy and robustness make it a compelling choice for applications that require high precision, such as social media moderation, compliance monitoring, and e-learning content curation. The dimension-expanding MLP architecture is shown in the figure 2 where hidden dimension are set to be always bigger than the input dimension.

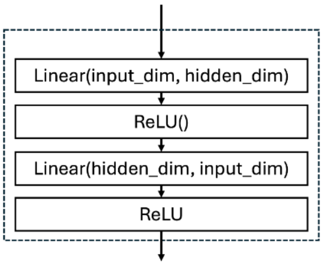


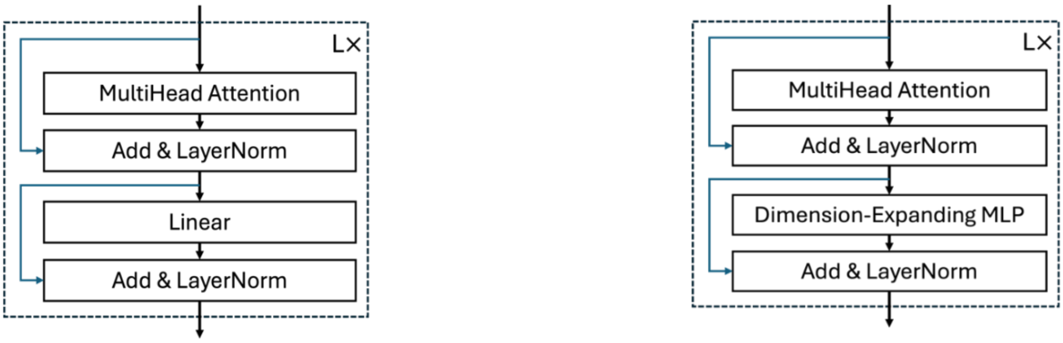
Figure 2. Dimension-Expanding MLP

3.3. Transformer Encoder

Following the original Transformer Networks, we build the Transformer Encoder with MultiHead Attention layer, add and normalize (skip connection and normalization), and Linear layer (fully connected layer) [19].

3.4. Transformer Encoder with Dimension-Expanding Multi-Layer Perceptron

In this study, we applied Dimension-Expanding MLP inside the Transformer Encoder as an alternative to the one linear layer as shown in figure 3. Overall, our proposed method is illustrated in figure 4.



Transformer Encoder one Linear layer (baseline)

Transformer Encoder with dimension-expanding MLP

Figure 3. Transformer Encoder

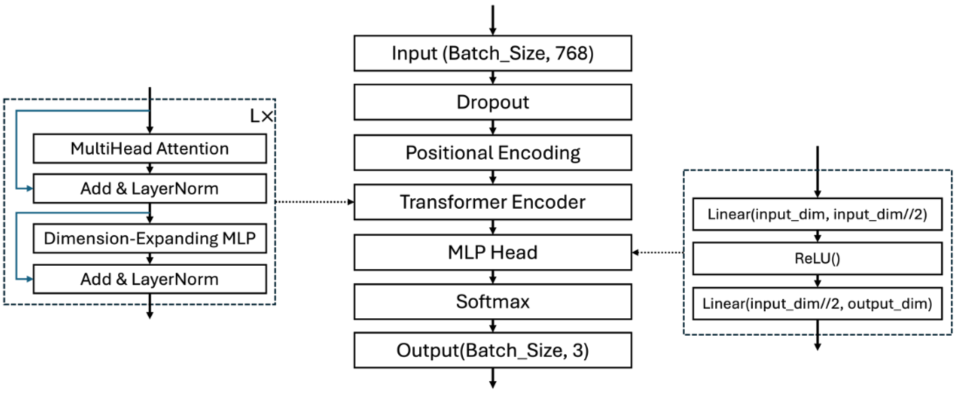


Figure 4. Dimension Expanded MLP in Transformer

3.5. Loss Function and Evaluation Metric

In this research, we used cross-entropy loss in the training to evaluate the learning of the model. Equation 2 shows the cross-entropy function implemented in our research.

$$\text{Loss} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C} y_{n,c}\log(\hat{y}_{n,c}) \qquad (2)$$

N is the number of the sample, C is the number of class, $y_{n,c}$ and $\hat{y}_{n,c}$ are the ground truth and prediction with respect to the sample n and class c. Meanwhile, for the testing scenario, we evaluate the model by confusion matrix (shown in table 1) and its variance of commonly used evaluation metrics for classification task. Table 1 shows the confusion matrix for the classification of the sentence/paragraph inappropriate content model. Eq. 3, 4, 5, and 6 are the metrics used to evaluate the testing data. In addition, we evaluate the model by the size of parameter and floating operation (floats) to show the inference without dependence to the device used for the fair comparison.

$$\text{Accuracy} = \frac{\text{TP+FN}}{\text{TP+FN+TN+FP}} \tag{3}$$

Table 1. Confusion matrix for sentence/paragraph inappropriate content classification

Actual / Predicted	Normal	Less Sensitive	Sensitive
Normal	TP	FP	FP
Less Sensitive	FN	TP	FP
Sensitive	FN	FN	TP

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FN}} \tag{4}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FP}} \tag{5}$$

$$F1\ \text{Score} = 2 \times \frac{\text{Precision}\times\text{Recall}}{\text{Precision+Recall}} \tag{6}$$

3.6. Experiment Setup

This research is executed using a Python PyTorch environment on a GPU NVidia GeForce RTX 4090 and an AMD RYZEN 7. The input features are bundled into a batch size of 128 with an embedding size of 768.

4. Results and Discussion

In this research, several experiments were carried out to test the performance of our proposed transformer in the cases of inappropriate sentences and paragraph classification for the filtering of digital content. There are 3 subsections to demonstrate the results.

4.1. Model Comparison to Baseline

At this phase, we evaluate the performance of our proposed model, Dimension-Expanding MLP Transformer, against the baseline model, one linear layer Transformer. The evaluation centers on essential metrics including accuracy, parameter count, and computational complexity in terms of FLOPs (Floating Point Operations per Second). Parameter count refers to the number of learnable weights in the model, which directly impacts the model's capacity to learn complex patterns. A higher parameter count generally means the model can handle more intricate tasks but may require more memory and processing power. FLOPs (Floating Point Operations) measure the total number of mathematical computations the model performs during inference. This metric provides insight into the computational cost of running the model. For filtering tasks, a lower FLOP count indicates faster processing, which is critical for real-time applications like social media moderation or live chat filtering, where speed and efficiency are essential. Our objective is to assess the balance between performance and resource utilization across different configurations of hidden dimensions, layers, and dropout rates (see table 2 below).

From the results presented in table 2, notably, the dimension-expanding MLP Transformer generally outperforms the one Linear layer Transformer in terms of accuracy across various settings. With a dropout rate of 0.2 and 2 layers, the Dimension-Expanding MLP Transformer achieves an accuracy rate of 0.737, which is slightly higher than the 1 Layer MLP Transformer's 0.7285. The dimension-expanding MLP Transformer's highest accuracy, 0.7445, occurs with a dropout of 0.3 across 2 layers, outperforming all other model configurations. Nevertheless, the one linear layer Transformer surpasses the dimension-expanding MLP Transformer in certain settings, like with 3 layers and a dropout of 0.2, where it scores 0.7425 compared to dimension-expanding multi-layer perceptron's 0.728. In terms of complexity, the dimension-expanding MLP Transformer tends to be more intricate than the one linear layer Transformer. For instance, as indicated in table 2, when the model is configured with 3 layers and a 0.1 dropout, the dimension-expanding MLP Transformer operates with 26.5859 million parameters, slightly higher than the 25.6019 million parameters in the one linear layer Transformer. This increase in parameters directly influences the number of operations required during both the forward and backward propagation stages. Additionally, the architectural complexity of the dimension-expanding MLP Transformer plays a critical role. Unlike the one linear layer Transformer, the dimension-expanding MLP introduces additional computational steps to process the expanded dimensions, which increases the overall computational load. This is evident in the GFLOPs comparison: for a configuration with 5 layers and a 0.1 dropout, the dimension-expanding MLP Transformer requires 645.181 GFLOPs, significantly higher than the one linear layer Transformer's 484.036 GFLOPs. In summary, the higher FLOP count of the dimension-expanding MLP Transformer is the result of a combination of its higher parameter count and the more complex operations stemming from its architectural design. While this added complexity contributes to improved accuracy (e.g., achieving a maximum accuracy of 0.7445 with a dropout of 0.3), it also makes the model more computationally demanding, underscoring the trade-off between accuracy and computational efficiency.

Table 2. Transformers Comparison

hidden dim	n layers	dropout	Ours			1 Layer Linear		
			accuracy	parameters	FLOPs	accuracy	parameters	FLOPs
1024	1	0.1	0.7405	15.5580 M	129.097G	0.7360	14.5739 M	96.867G
1024	2	0.1	0.7395	21.0719 M	258.118G	0.7135	20.0879 M	193.660G
1024	3	0.1	0.7285	26.5859 M	387.139G	0.6810	25.6019 M	290.452G
1024	4	0.1	0.6570	32.0999 M	516.160G	0.7195	31.1158 M	387.244G
1024	5	0.1	0.7230	37.6139 M	645.181G	0.6685	36.6298 M	484.036G
1024	1	0.2	0.7440	15.5580 M	129.097G	0.7380	14.5739 M	96.867G
1024	2	0.2	0.7370	21.0719 M	258.118G	0.7285	20.0879 M	193.660G
1024	3	0.2	0.7280	26.5859 M	387.139G	0.7425	25.6019 M	290.452G
1024	4	0.2	0.6880	32.0999 M	516.160G	0.7395	31.1158 M	387.244G
1024	5	0.2	0.7080	37.6139 M	645.181G	0.7185	36.6298 M	484.036G
1024	1	0.3	0.7395	15.5580 M	129.097G	0.7370	14.5739 M	96.867G
1024	2	0.3	0.7445	21.0719 M	258.118G	0.7410	20.0879 M	193.660G
1024	3	0.3	0.7210	26.5859 M	387.139G	0.7110	25.6019 M	290.452G
1024	4	0.3	0.7210	32.0999 M	516.160G	0.7205	31.1158 M	387.244G
1024	5	0.3	0.7305	37.6139 M	645.181G	0.7210	36.6298 M	484.036G

4.2. Model Comparison to RNN-Based Method

In this research, we perform a comparison study to evaluate the model. This stage examines multiple scenarios with hyperparameters such as the number of layers and the probability of dropout to evaluate the performance of the GRU, LSTM, and Transformer models. The model settings consist of a hidden dimension set to 1024, a number of layers between 1 and 5, dropout rates of 0.1, 0.2, and 0.3, along with a learning rate of 0.001, resulting in a total of 15 different scenarios. In other way, we set different optimizers, GRU and LSTM used AdamW, while Transformer using SGD, the reason is because Transformer uses an attention mechanism architecture, so it is very suitable for utilizing SGD because gradual optimization of Transformer parameters can support small adjustments to gradients. This approach is effective in maintaining model stability during training, especially on large datasets. Meanwhile, GRU and LSTM use a more complex internal memory structure (gated mechanism), so they require more adaptive optimization, such as AdamW. The results can be seen in table 3 below.

According to table 3, LSTM achieved its highest accuracy of 0.744 on the 14th trial, Transformer reached its maximum accuracy of 0.7445 on the 12th trial, and GRU achieved its highest accuracy of 0.739 on the 8th trial. Our model with a dropout rate of 0.3 demonstrated superior performance, achieving the highest accuracy of 0.7445. This outperformed the baseline Transformer model, which achieved an accuracy of 0.7425 under the same dropout condition, and traditional methods like LSTM and GRU, which achieved maximum accuracies of 0.744 and 0.739, respectively. Regarding parameter count, the LSTM model occasionally has the fewest parameters in certain trials, such as in the initial trial with a mere 7.35M. In contrast, the Ours model has more parameters (15.56M), while GRU has 5.51M parameters. In particular, in the trial with the maximum parameter count (40.94M for LSTM), the Ours model still features more parameters than LSTM, but remains slightly below GRU. In terms of FLOPs, the LSTM model consistently exceeds GRU. For example, in the first trial, LSTM achieved 1.44 TFLOPs and GRU only 1.08 TFLOPs. Despite having more parameters, the Ours model demonstrates enhanced computational efficiency in terms of GFLOPs, reporting only 129.1 GFLOPs in the first trial, significantly lower than LSTM's FLOPs. However, as the parameter count increases to 40.94M, the FLOPs of the Ours model increase to 645.2 GFLOPs, exceeding GRU but still lower than LSTM. Overall, the Ours model strikes a balance between parameter size and computational efficiency. It includes more parameters than LSTM but achieves lower FLOPs, indicating strong efficiency. For a detailed analysis, we have visualized the decrease in loss and confusion matrix for the best model. These visualizations are presented in figure 5 and figure 6.

Table 3. Hyperparameters and Model Comparison

n **layers**	**dropout**	**Accuracy↑**			**Parameters↓**			**FLops↓**		
		LSTM	**Ours**	**GRU**	**LSTM**	**Ours**	**GRU**	**LSTM**	**Ours**	**GRU**
1	0.1	0.7350	0.7405	0.7265	7.3513M	15.5580M	5.5142M	1.4441T	129.097G	1.083T
2	0.1	0.7365	0.7395	0.7315	15.7481M	21.0719M	11.8118M	3.0891T	258.095G	2.322T
3	0.1	0.7330	0.7285	0.7145	24.1449M	26.5859M	18.1094M	4.7359T	387.092G	3.563T
4	0.1	0.7335	0.6570	0.7270	32.5417M	32.0999M	24.4070M	6.3893T	516.189G	4.803T
5	0.1	0.7250	0.7230	0.7285	40.9385M	37.6139M	30.7046M	8.0432T	645.236G	6.031T
1	0.2	0.7285	0.7440	0.7210	24.1449M	15.5580M	5.5142M	4.7359T	387.092G	3.563T
2	0.2	0.7360	0.7370	0.7210	7.3513 M	21.0719M	11.8118M	3.0891T	258.095G	2.322T
3	0.2	0.7345	0.7280	0.7390	15.7481M	26.5859M	18.1094M	4.7359T	387.092G	3.563T
4	0.2	0.7395	0.6880	0.7385	24.1449M	32.0999M	24.4070M	6.3893T	516.189G	4.803T
5	0.2	0.7195	0.7080	0.7360	32.5417M	37.6139M	30.7046M	8.0432T	645.236G	6.031T
1	0.3	0.7395	0.7395	0.7185	40.9385M	15.5580M	5.5142M	8.0432T	645.236G	6.031T
2	0.3	0.7365	0.7445	0.7300	7.3513 M	21.0719M	11.8118M	3.0891T	258.095G	2.322T
3	0.3	0.7330	0.7210	0.7215	15.7481M	26.5859M	18.1094M	4.7359T	387.092G	3.563T
4	0.3	0.7440	0.7210	0.7370	24.1449M	32.0999M	24.4070M	6.3893T	516.189G	4.803T
5	0.3	0.7195	0.7305	0.7290	32.5417M	37.6139M	30.7046M	8.0432T	645.236G	6.031T

The loss graph in figure 5 illustrates a general decrease in loss values for both the training and validation datasets, suggesting effective model learning. By the end of the training process, the training loss stabilizes around 0.80, while the validation loss fluctuates within a range of 0.80 to 0.90. These values indicate that while the model successfully minimizes errors on the training set, the fluctuations in validation loss highlight potential challenges in generalization. Including these specific values provides a clearer understanding of the graph's significance and the model's performance trends.

In the confusion matrix figure 6, the performance of the model varies between the three classes. Class 0 demonstrates the highest accuracy with 1031 correct predictions, while class 1 performs the poorest, with only 194 correct predictions and frequent misclassifications, predominantly in class 0. Class 2's performance is intermediate, achieving 264 correct predictions, but still experiencing errors classified as both class 0 and class 1. In summary, the model performs effectively on the majority class (class 0) but struggles to accurately differentiate the minority classes (classes 1 and 2).

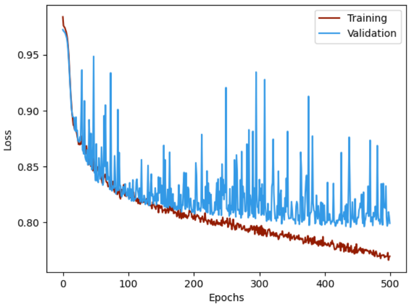


Figure 5. Training loss evaluation on 500 epochs.

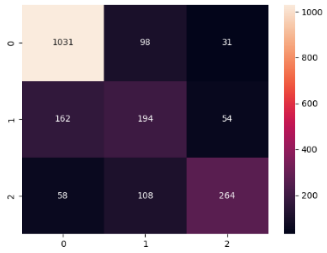


Figure 6. Confusion matrix with dimension-expanding MLP Transformer.

4.3. Use Cases and Applications

Our models inherently detect, screen, and manage unsuitable content. As digital platforms expand in size and influence, prioritizing the security, professionalism, and suitability of user-generated content has become essential. Whether observing social media interactions or ensuring legal documents meet standards, these models offer strong solutions across diverse sectors. Use cases that underscore both the extensive relevance and significance of these technologies in today's digital realms. Each sector shows how organizations can deploy models to automate routine moderation tasks, enhance user experiences, and adhere to sector-specific regulations. By customizing these models to their specific requirements, companies can create safer, more interactive, and compliant digital environments for their users and partners.

Content moderation technology is used on various platforms to detect and filter offensive language, harassment, and policy violations in real time. Ensure compliance in e-learning, online marketplaces, news platforms, and workplace tools by screening user-generated content for inappropriate or misleading information. Additionally, it aids legal and compliance sectors by identifying non-compliant contract clauses, while also enhancing safety in gaming, media entertainment, and customer support by filtering abusive messages. Parental control apps, healthcare platforms, and government monitoring tools utilize this technology to protect minors, detect distress signals, and prevent cyber threats. Lastly, it supports recruitment platforms and custom enterprise solutions by maintaining professionalism and ensuring the content compliance with business policies.

5. Conclusion

This study presents a novel method for filtering inappropriate digital content through the integration of dimension-expanding multi-layer perceptron into transformer architectures. The introduced model adeptly merges the advantages of dimension-expanding multi-layer perceptron, namely efficient data processing and improved feature refinement, with the capability of Transformers to capture both global and local contexts. The research emphasizes the practical applications of the model in areas such as social media content moderation, legal document compliance monitoring, and filtering harmful content on e-learning and gaming platforms. By automating intricate tasks, the model helps create safer, more ethical, and engaging digital environments. However, future research could explore several areas for improvement to further enhance the model's capabilities. One key direction is the integration of multilingual support, allowing the model to process and filter content in a wider range of languages while maintaining accuracy and contextual understanding. This would be particularly valuable for global platforms with diverse user bases. Additionally, future work could focus on testing the model on larger, real-world datasets that reflect the complexities of live environments, such as social media platforms or gaming communities, to validate its robustness and adaptability. Another promising area is the incorporation of sparsity mechanisms or model pruning techniques to reduce computational requirements without sacrificing performance, making the model more efficient for deployment in resource-constrained environments. Furthermore, enhancements to contextual embedding strategies, such as leveraging domain-specific pretraining, could improve the model's sensitivity to nuanced language patterns, particularly in identifying subtle inappropriate content.

6. Declarations

6.1. Author Contributions

Conceptualization: A.C.W., A.P.Y., R.A., M.A.L., and M.S.; Methodology: M.S.; Software: A.C.W.; Validation: A.C.W., M.S., and R.A.; Formal Analysis: A.C.W., M.S., and R.A.; Investigation: A.C.W.; Resources: M.S.; Data Curation: M.S.; Writing Original Draft Preparation: A.C.W., M.S., and R.A.; Writing Review and Editing: M.S., A.C.W., and R.A.; Visualization: A.C.W. All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

This study is funded by DRTPM from the Ministry of Education, Culture, Research and Technology (Kementrian Pendidikan, Kebudayaan, Riset, dan Teknologi) Republic of Indonesia through the fundamental research grant 2024 (No. 108/E5/PG.02.00.PL/2024), as well as Institut Teknologi Telkom Purwokerto for providing facilities for conducting research before merging with Telkom University. Collaborating with the KBM app or KBM.id (PT. Sembilan Sembilan Media), a partner that provides data research.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] S. R. Hill, I. Troshani, and D. Chandrasekar, "Signalling Effects of Vlogger Popularity on Online Consumers," *Journal of Computer Information Systems*, vol. 60, no. 1, pp. 76–84, Jan. 2020.

[2] D. Bhatia, N. R. Sharma, J. Singh, and R. S. Kanwar, "Biological methods for textile dye removal from wastewater: A review," *Critical Reviews in Environmental Science and Technology*, vol. 47, no. 19, pp. 1836–1876, Oct. 2017.

[3] H. Chen, "Antecedents of positive self-disclosure online: an empirical study of US college students' Facebook usage," *Psychology Research and Behavior Management*, vol. 10, no. 42, pp. 147–153, May 2017.

[4] M. Frobenius, "Beginning a monologue: The opening sequence of video blogs," *Journal of Pragmatics*, vol. 43, no. 3, pp. 814–827, Feb. 2011.

[5] J. E. Lee and B. Watkins, "YouTube vloggers' influence on consumer luxury brand perceptions and intentions," *Journal of Business Research*, vol. 69, no. 12, pp. 5753–5760, Dec. 2016.

[6] D. Horton and A. Strauss, "Interaction in Audience-Participation Shows," *American Journal of Sociology*, vol. 62, no. 6, pp. 579–587, May 1957.

[7] G. S. Stever and K. E. Lawson, "Twitter as a way for celebrities to communicate with fans: Implications for the study of parasocial interaction," *North American Journal of Psychology*, vol. 15, no. 2, pp. 339–354, 2013.

[8] R. B. Rubin and M. P. McHugh, "Development of parasocial interaction relationships," *Journal of Broadcasting & Electronic Media*, vol. 31, no. 3, pp. 279–292, Jun. 1987.

[9] K. Sokolova and H. Kefi, "Instagram and YouTube bloggers promote it, why should I buy? How credibility and parasocial interaction influence purchase intentions," *Journal of Retailing and Consumer Services*, vol. 53, no. 2, pp. 1–9, Mar. 2020.

[10] K. Sokolova and C. Perez, “You follow fitness influencers on YouTube. But do you actually exercise? How parasocial relationships, and watching fitness influencers, relate to intentions to exercise,” *Journal of Retailing and Consumer Services*, vol. 58, no. 1, pp. 1–11, Jan. 2021.

[11] T. Hartmann and C. Goldhoorn, “Horton and Wohl Revisited: Exploring Viewers’ Experience of Parasocial Interaction,” *Journal of Communication*, vol. 61, no. 6, pp. 1104–1121, Dec. 2011.

[12] J. L. Dibble, T. Hartmann, and S. F. Rosaen, “Parasocial Interaction and Parasocial Relationship: Conceptual Clarification and a Critical Assessment of Measures,” *Human Communication Research*, vol. 42, no. 1, pp. 21–44, Jan. 2016.

[13] E. Katz, J. G. Blumler, and M. Gurevitch, “Uses and Gratifications Research,” *The Public Opinion Quarterly*, Winter, 1973.

[14] J. Raacke and J. Bonds-Raacke, “MySpace and Facebook: Applying the Uses and Gratifications Theory to Exploring Friend-Networking Sites,” *CyberPsychology & Behavior*, vol. 11, no. 2, pp. 169–174, Apr. 2008.

[15] H. Lim and A. Kumar, “Variations in consumers’ use of brand online social networking,” *Journal of Retailing and Consumer Services*, vol. 51, no. 6, pp. 450–457, Nov. 2019.

[16] M. S. B. C. B. C. & S. M. Kruse R., “Multi-layer perceptrons. In Computational intelligence: a methodological introduction,” *Cham: Springer International Publishing*, pp. 53–124, 2022.

[17] L. C. Y. Y. Z. J. & G. M. Zhang J., “Applications of artificial neural networks in microorganism image analysis,” *Artificial Intelligence Review*, vol. 56, no. 2, pp. 1013–1070, 2023.

[18] W. G. Z. C. P. P. C. R. F. C. S. A. S. V. D. L. Liu C. and H. Wu, “End-to-end methane gas detection algorithm based on transformer and multi-layer perceptron,” *Optics Express*, vol. 32, no. 1, pp. 987–1002, 2023.

[19] S. N. M. P. N. U. J. J. L. G. A. N. K. L. & P. I. Vaswani A., “Attention is All You Need,” *Neural Information Processing Systems*, 2017.

[20] G. Mahesh and R. Mittal, “Digital content creation and copyright issues,” *Electronic Library*, vol. 27, no. 4, pp. 676–683, 2009.

[21] E. Risdianto and E. Apiri, “Analysis of the Implementation of Project-Based Learning Models,” *JENTIK: Jurnal Pendidikan Teknologi Informasi dan Komunikasi*, vol. 1, no. 1, pp. 6–12, 2022.

[22] E. López-Meneses, F. M. Sirignano, E. Vázquez-Cano, and J. M. Ramírez-Hurtado, “University Students’ Digital Competence,” *Australasian Journal of Educational Technology*, vol. 36, no. 3, pp. 69–88, 2020.

[23] N. Kalajdžisalihović, L. Kasumagić-Kafedžić, and A. Sadiković, “Digital Literacy, Digital Pedagogy, and Digital Content Creation,” *Educational Role of Language Journal*, vol. 8, no. 2, pp. 82–90, 2023.

[24] M. L. B. dos Santos, “The ‘so-called’ UGC: An Updated Definition of User-Generated Content in the Age of Social Media,” *Online Information Review*, vol. 45, no. 1, pp. 2–19, 2021.

[25] H. Cowie, “Cyberbullying and its impact on young people’s emotional health and well-being,” *The Psychiatrist*, vol. 37, no. 5, pp. 167–170, 2013.

[26] P. K. Smith, J. Mahdavi, M. Carvalho, S. Fisher, S. Russell, and N. Tippett, “Cyberbullying: Its Nature and Impact in Secondary School Pupils,” *Journal of Child Psychology and Psychiatry*, vol. 49, no. 4, pp. 376–385, 2008.

[27] R. S. Tokunaga, “Following You Home: A Critical Review and Synthesis of Research on Cyberbullying Victimization,” *Computers in Human Behavior*, vol. 26, no. 3, pp. 277–287, 2010.

[28] B. Garner, “Teaching Students to Become Digital Content Curators: Fact or Fiction?,” *Cambridge Scholars Publishing*, 2019.

[29] C. Rus-Casas, D. Eliche-Quesada, F. J. Muñoz-Rodríguez, and M. D. La Rubia, “Content Curation in E-Learning: A Case of Study with Spanish Engineering Students,” *Applied Sciences*, vol. 12, no. 6, p. 3188, 2022.

[30] H. L. Rhee, “A new lifecycle model enabling optimal digital curation,” *Journal of Librarianship and Information Science*, vol. 56, no. 1, pp. 241–266, 2022.

[31] D. Liao, “Sentence Embeddings using Supervised Contrastive Learning,” *CoRR*, vol. abs/2106.04791, 2021.

[32] M. A. Rahman, “Impact of Transformer-Based Models in NLP: An In-Depth Study on BERT and GPT,” *IEEE Xplore*, 2023.

[33] M. A. Rahman, “Vision Transformers for Image Classification: A Comparative Survey,” *Technologies*, vol. 13, no. 1, p. 32, 2023.