ISSN 2301-8984 (Print), 2809-1884 (Online)
Published by Institute of Computer Science

Implementation of vision transformer for offensive language detection on tiktok social media

Zulekha Rahmawaty[1], Fatsyarina Fitriastuti[2], Ryan Ari Setyawan[3]
[1,2,3]Informatics Engineering, Faculty of Engineering, Universitas Janabadra Yogyakarta, Indonesia

ARTICLE INFO

Article history:

Received Jul 1, 2025
Revised Jul 10, 2025
Accepted Jul 19, 2025

Keywords:

Content Moderation;
Deep Learning;
Offensive Language Detection;
Tiktok;
Vision Transformer.

ABSTRACT

The rise of social media platforms such as TikTok has introduced new challenges in content moderation, particularly concerning the spread of offensive language and hate speech. One promising approach to addressing this issue is through automatic detection using deep learning technology. This study implements the Vision Transformer (ViT) to detect offensive language on the TikTok platform based on visual data in the form of comment screenshots. The dataset used consists of 1,401 labeled images categorized into two classes: offensive and non-offensive. The training process was conducted over 50 epochs without a validation split, and the evaluation was carried out using accuracy, precision, recall, and F1-score metrics. Results showed high performance, with an accuracy of 99.93%, precision of 0.9979, recall of 1.000, and F1-score of 1.000 at the 40th epoch, maintaining stability through the end of training. These findings demonstrate that ViT is effective in extracting visual features from image-based comments, even without access to raw text. This approach is particularly relevant in the context of TikTok, where comments often appear in visual formats such as thumbnails, screenshots, or reaction videos. This research opens up opportunities for the implementation of image-based offensive language detection systems that can enhance content moderation by adapting to various visual formats. Further development is recommended using a larger dataset and more systematic data splitting to test the model's generalization capability.

Corresponding Author:

Zulekha Rahmawaty,
Department of Informatics Engineering,
Universitas Janabadra Yogyakarta,
Jl. Tentara Rakyat Mataram No.55-57, Bumijo, Kec. Jetis, Kota Yogyakarta, Daerah istimewa Yogyakarta 55231, Indonesia
Email: zulekharahmawaty078@gmail.com

1. INTRODUCTION

The rapid development of information and communication technology has driven the growth of social media as one of the main platforms for interaction and information sharing. One of the most prominent platforms is TikTok, a short video-based social media platform that has gained global popularity and serves as a space for creative expression among its users. However, behind its popularity, TikTok also presents serious challenges, particularly in the spread of offensive language, hate speech, and other negative content. This phenomenon has a significant impact on mental health, especially among vulnerable groups such as children and adolescents. Jadmiko & Damariswara. (2022) revealed that teenagers tend to imitate offensive language they encounter on TikTok, which can potentially disrupt their social and emotional development. Research by Bilali et al. (2025) also showed that problematic use of TikTok is closely associated with increased symptoms of anxiety, depression, and excessive sleepiness among adolescents, with a greater impact observed in male users. In addition, Conte et al. (2025), through a review of 20 studies involving 17,336 respondents from 10 countries, found that TikTok may decrease life satisfaction, increase the risk of behavioral addiction, influence body image and self-esteem, and contribute to the spread of certain mental disorder symptoms. While the platform offers opportunities for self-expression and social connection, these findings highlight the urgent need to address the psychological impacts of TikTok use, particularly among adolescents who are still in the process of emotional and identity development.

On social media platforms like TikTok, offensive comments not only disrupt user comfort but can also impact mental health and foster an unhealthy communication culture. A study by Jadmiko & Damariswara. (2022) found that teenagers often imitate offensive language encountered on TikTok, making this phenomenon a real threat to the social and emotional development of younger generations. Research by Nabila Budihartana & Sudrajat (2024) also highlights the potential negative implications for communication ethics among users, particularly adolescents. Therefore, the automatic detection of offensive language on social media has become an urgent necessity as a strategic step to strengthen content moderation.

Various Natural Language Processing (NLP)-based approaches have been developed to detect offensive comments. One commonly used method is the BERT (Bidirectional Encoder Representations from Transformers) model, which is capable of deeply understanding sentence context (Rendragraha et al., 2021) However, text-based approaches have limitations, particularly on platforms like TikTok, where comments often appear in visual formats such as screenshots or thumbnails rather than raw text.

Since many comments are conveyed in image format, computer vision-based approaches are becoming increasingly relevant. Research by Sulistiyawati et al. (2022) highlights the need for the development of visual-based automatic detection methods to identify offensive communication. The Vision Transformer (ViT) has gained attention as an effective deep learning method for processing image data by modeling spatial and contextual relationships simultaneously.As stated by Islam. (2022) ViT has successfully addressed various visual problems by focusing on long-range relationships, which further supports its advantage over traditional models. ViT processes images similarly to how text tokens are handled, making it highly effective in capturing offensive patterns in visual content (Khan et al., 2022).

The Vision Transformer (ViT) model, introduced by Dosovitskiy et al. in 2021, has proven to be highly effective in image classification tasks. The image is segmented into smaller patches, which are then treated as sequential inputs similar to tokens in natural language processing,ViT is capable of capturing a global understanding of the image (Romindo et al., 2023).This approach is particularly well-suited for detecting offensive comments presented in image format on TikTok, especially in the form of screenshots or video content containing abusive text.

This study aims to implement the Vision Transformer (ViT) model to detect offensive language on the TikTok social media platform using a dataset of 1,401 manually labeled comment screenshots. Labeling using Roboflow is a popular method for preparing image datasets for ViT model training. The training process was conducted over 50 epochs, with evaluation using metrics such as accuracy, precision, recall, and F1-score, which are widely implemented for assessing binary classifier performance (Contreras Ortiz et al., 2025). After training, these metrics were stored in a metrics.json file and visualized using Matplotlib, following standard practices in ViT-based model training and other image classification tasks.

This research is expected to contribute to the development of image-based automated content moderation systems on social media platforms. Moreover, this approach demonstrates significant potential in utilizing visual data to detect offensive content, and it paves the way for the integration of multimodal approaches combining text and image to produce more adaptive and accurate detection systems (Romindo et al., 2023).

Despite the advancements in offensive content detection using image-based approaches, one of the critical challenges in applying such methods on platforms like TikTok lies in the limited access to APIs and restrictions in retrieving raw comment data. Unlike platforms that allow direct access to structured text through APIs, TikTok often presents user interactions in image or video formats, complicating the data extraction process. This constraint necessitates alternative solutions, such as screenshot-based datasets, which although effective, introduce additional preprocessing steps and may limit the scalability of the moderation system. Therefore, the choice to adopt a computer vision-based method is not only driven by data characteristics but also by platform-specific limitations in data accessibility.

2. RESEARCH METHOD

This research falls under the category of quantitative experimental study, as it examines the performance of a deep learning model based on Vision Transformer (ViT) in detecting offensive language through visual data from TikTok. Quantitative research fundamentally utilizes numerical data to systematically and objectively analyze phenomena. Kurniawan & Puspitaningtyas (2023) explain that this approach emphasizes the measurement of variables in quantifiable statistical terms. In this study, the model is not only evaluated but also developed and trained through an experimental process using a structured dataset.

The nature of this research is descriptive, focusing on illustrating the performance of the ViT model in classifying images into two categories: offensive and non-offensive comments. According to (Jannah et al., 2022) quantitative descriptive research aims to explain a phenomenon through patterns revealed in the collected quantitative data. Model evaluation is conducted using metrics such as accuracy, precision, recall, and F1-score, which are analyzed from the training results over 50 epochs. This approach allows researchers to observe how well the model can recognize patterns from visual data

The quantitative approach in this study was chosen because all stages from preprocessing to model performance evaluation were carried out using numerical and statistical calculations. Sujarweni & Wiratna (2023) emphasize that the quantitative method allows for data generalization and objective analysis, making it suitable for research involving large-scale data processing or statistical modeling. The training results are presented in the form of graphs to illustrate the model's performance during training and to provide a visual overview of the model's learning process.

The initial stage of this research was data collection. The data were obtained from the TubeTrendy platform, which archives TikTok videos and their user comments. The collected data consisted of screenshots of user comments. The decision to use visual data was based on the fact that comments on TikTok are often not available in extractable text format. This aligns with (Abdullah, (2015) perspective that data collection in quantitative research is not limited to traditional numerical data but can also include digital objects that are converted into a processable format.

By using visual data, researchers can fully capture the characteristics of the comments, including their appearance, color, and text formatting. This allows for a deeper analysis of the visual aspects of abusive comments. Comments visualized through screenshots also reflect the real use of TikTok, where much of the content appears in the form of interface images. This strategy expands the scope of modern quantitative research, as stated by Jannah et al.(2022), who argue that this method can be combined with technological approaches to extract deeper insights.

The collected data was then manually labeled using the Roboflow platform, which was selected for its efficient image dataset management capabilities. The labeling process resulted in two categories: abusive language and non-abusive language. Of the 1,401 available images, 933 were categorized as abusive comments, and 468 as non-abusive. Due to the limited amount of data, all samples were used for training without a validation split. Although this approach does not assess the model’s generalization to unseen data, it is sufficient for an initial experimental stage and for monitoring the model’s performance trends. However, it is worth reflecting to what extent the dataset size (1,401 images) is sufficient to support claims of model stability, especially within the scope of a visual-based binary classification task.

The next stage was image preprocessing. The images were resized to 224x224 pixels to match the input requirements of the Vision Transformer. This transformation was performed using the torchvision. transforms library within the PyTorch pipeline. The images were also converted into tensors to be processed by the neural network. This procedure aligns with recommendations in technology-based quantitative research, where input data standardization is a critical prerequisite for achieving accurate and reliable training results (A. W. Kurniawan & Puspitaningtyas, 2023).

Research Flow

This research began with the process of problem identification and objective formulation, namely to develop an image-based abusive comment detection model on the TikTok platform using the Vision Transformer (ViT). The next step involved data collection in the form of screenshot images (thumbnails) of TikTok comments from the TubeTrendy platform, which provides a collection of public comments in visual format.

The collected data was then manually labeled using the Roboflow platform, categorizing the images into two groups: abusive and non-abusive comments. Labeling was carried out carefully to ensure the quality of the training data. The next stage was image preprocessing, which included resizing the images to 224×224 pixels and converting them into tensors to be compatible with the ViT architecture. The Vision Transformer (ViT) model with the vit_base_patch16_224 architecture was then initialized and adjusted to match the number of classes.

The training process was conducted gradually over 50 epochs using the Adam optimizer and the CrossEntropyLoss function. At each epoch, the model generated predictions that were compared to the ground-truth labels to calculate evaluation metrics such as accuracy, precision, recall, and F1-score. These metrics were stored in a JSON file and the training log, and then visualized as graphs to analyze the model's performance trends over time. The overall research flow is illustrated in Figure 1 below.

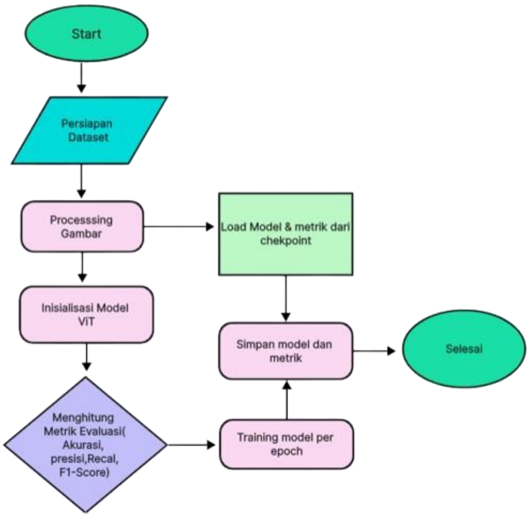

Figure 1. Research flow

Model Architecture and Training Configuration

The model used in this study is the Vision Transformer (ViT), which was chosen for its ability to capture spatial and contextual representations of images more effectively through the self-attention mechanism, as well as its competitive performance compared to traditional convolutional models in various image classification tasks (Zhang et al., 2023).

The architecture used in this study is vit_base_patch16_224 from the timm library, which is an implementation of the Vision Transformer (ViT) model introduced by Dosovitskiy et al (2021).This architecture adapts the Transformer approach for computer vision tasks by dividing the input image into small patches of 16×16 pixels, then processing them as a sequence of visual tokens, similar to how words are processed in Natural Language Processing (NLP).

The model was initialized using pretrained weights from the ImageNet dataset (transfer learning technique), and the classification head was modified to match the number of output classes (2 outputs). This adjustment is essential because the pretrained ViT was originally trained for 1,000-class classification tasks.

The model training process was carried out with the following configuration: a) Optimizer: Adam, nown for its fast convergence and suitability for deep learning classification tasks; b) Learning rate: 0.0001, selected to maintain stability during training; c) Loss function: CrossEntropyLoss, commonly used for multi-class classification; d) Number of epochs: Trained incrementally up to 50 epochs; e) Checkpointing: Model disimpan The model was saved at the end of each epoch to allow training to be resumed in case of interruptions (e.g., Colab timeout or power outage).

Evaluation Metrics and Logging

The model was evaluated using four main classification metrics: a) Accuracy: Persentase The percentage of correct predictions out of all predictions; b) Precision: the correctness of positive predictions, i.e., how many of the predicted positives are truly positive; c) Recall: The model's ability to capture actual positive data; e) F1-Score:The harmonic mean of precision and recall, which serves as an overall performance indicator, especially when there is class imbalance.

All metric values were stored in a metrics.json file, while the training log for each epoch was recorded in log_training.txt. The final results were visualized using the Matplotlib library in the form of performance graphs per epoch, showing the progress of accuracy, precision, recall, and F1-score from epoch to epoch. This visualization served as the foundation for analyzing the model's performance in the *Result and Discussion* section.

Evaluation Formulas

To evaluate the performance of the Vision Transformer (ViT) model in classifying TikTok comments into two classes (offensive and non-offensive), four main evaluation metrics were used: accuracy, precision, recall, and F1-score. These metrics are essential to assess how well the model consistently and accurately distinguishes between the two classes. The following are the formulas and definitions:

1. *Accuracy,* measures the proportion of correct predictions out of the total data evaluated.

$$\text{Akurasi} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

2. Precision, measures the accuracy of positive predictions, i.e., how many of the comments predicted as “offensive” are truly offensive.

$$\text{Presisi} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

3. Recall (Sensitivity), measures how many offensive comments were correctly identified by the model.

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

4. F1-Score, the harmonic Mean of precision and recall, used to evaluate the balance between detection capability and prediction accuracy, especially when the data is imbalanced.

$$\text{F1} - \text{Score} = \frac{\text{Presisi x Recall}}{\text{Presisi} + \text{Recal}}$$

Explanation

- TP (True Positive): Offensive comments correctly classified as offensive.
- TN (True Negative): Non-offensive comments correctly classified as non-offensive.
- FP (False Positive): Non-offensive comments incorrectly classified as offensive.
- FN (False Negative): Offensive comments that were not detected by the model.

3. RESULTS AND DISCUSSIONS

Model Training Results

The training process of the Vision Transformer (ViT) model was carried out over 50 epochs in stages using a dataset consisting of screenshot images of TikTok comments that had been manually classified into two categories: offensive and non-offensive language. The dataset contained a total of 1,401 images, with 933 labeled as offensive and 468 as non-offensive

The model was trained without a validation split, meaning all data were used solely for training. This approach was taken due to the limited size of the dataset. However, it still allowed for monitoring the learning trends of the model throughout the training process by recording key evaluation metrics accuracy, precision, recall, and F1-score at each epoch.

The model was trained using the Adam optimizer with a learning rate of 0.0001 and the CrossEntropyLoss function, commonly used for multi-class classification tasks. To prevent loss of progress due to system interruptions (such as Google Colab timeouts), model weights were saved at the end of each epoch (checkpointing) to Google Drive.

Table 1. Performance metrics of vit model at selected epochs

	Training			
Epoch	Accuracy (%)	*Precision*	*Recall*	*F1- Score*
1	63,24	*0,5190*	0,5094	0,4814
5	76,52	*0,7420*	0,7065	0,7174

Training				
Epoch	Accuracy (%)	*Precision*	*Recall*	*F1- Score*
10	93,29	*0,9303*	0,9177	0,9235
15	99,00	*0,9898*	0,9877	0,9887
20	98,64	*0,9829*	0,9808	0,9818
25	97,72	*0,9619*	0,9701	0,9660
30	98,93	*0,9808*	0,9829	0,9840
35	98,50	*0,9806*	0,9744	0,9775
40	99,93	*0,9979*	1,000	1,000
45	99,36	*0,9895*	0,9915	0,9904
50	98,86	*0,9829*	0,9829	0,9829

From the table, the results indicate that the model experienced a significant performance improvement starting from the initial epochs. Beginning with an accuracy of 63.24% at epoch 1, the model successfully achieved over 99% accuracy after epoch 15. This improvement reflects the effectiveness of the Vision Transformer (ViT) architecture in learning visual patterns from TikTok comment data.

In addition to accuracy, both precision and recall also increased significantly. Precision rose from 0.5190 at epoch 1 to over 0.98 after epoch 15, indicating The capability of the model to accurately identify offensive comments with minimal false positive classifications. Similarly, recall improved from 0.5094 to above 0.98, showing that the model was able to recognize most offensive comments without missing many (low false negatives).

The F1-score, which represents the harmonic mean of precision and recall, also demonstrated consistent model performance, remaining above 0.98 from epoch 15 onwards. This stability confirms that the model is not only accurate but also balanced in handling both classes offensive and non-offensive comments effectively.

Visualization and Graph Analysis

To provide a more comprehensive understanding of the Vision Transformer (ViT) model's performance during training, visualizations of four key classification metrics Accuracy, Precision, Recall, and F1-Score were generated. These metrics were computed at each epoch and stored in the metrics.json file. Upon completion of training, the metric data were visualized in the form of graphs using the matplotlib library.

The graphs illustrate the model's performance progression from epoch 1 to epoch 50. In general, all metrics showed significant improvement as the number of epochs increased, with a trend approaching maximum values (near 1.0 or 100%) after epoch 15.

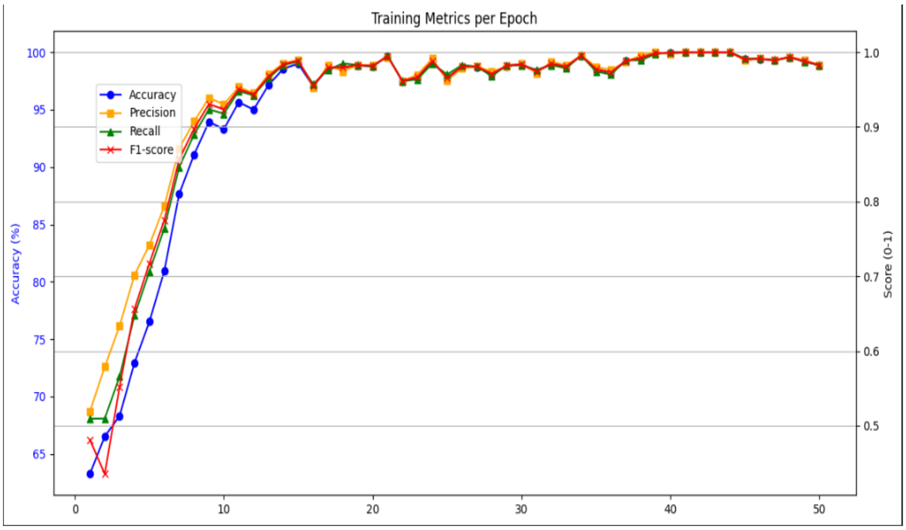

Figure 2. Training accuracy over epochs

The Vision Transformer (ViT) model used in this study demonstrated excellent performance in detecting toxic comments on the TikTok social media platform, as reflected by a significant improvement across Performance was analyzed based on key statistical indicators accuracy, precision, recall, and F1-score throughout the 50 training epochs. In the early phase of training, the model’s accuracy was relatively low, starting at 63.24% in the first epoch. However, the improvement occurred rapidly, and by epoch 15, accuracy had surpassed 99%. After this point, the model’s performance remained stable, with accuracy ranging between 98.5% and 99.9% until the end of training. This stability indicates that the model not only learned quickly but also consistently maintained a high classification capability throughout the training process.

Precision, as an indicator of the model’s accuracy in predicting the positive class (toxic comments), also showed a sharp upward trend. Initially, the precision score was 0.5190 in the first epoch, but it steadily increased to 0.9898 by epoch 15. This high precision indicates that the model rarely produced false positives.i.e., neutral comments were seldom misclassified as toxic. This is crucial in ensuring accurate content moderation and preventing the over-filtering of harmless comments.

In addition to precision, the model’s recall score also exhibited significant improvement. Starting from 0.5094 in the first epoch, recall increased to 0.9877 by epoch 15 and even reached a perfect score of 1.000 at epoch 40. A recall of 1.000 indicates that the model was able to identify nearly all toxic comments in the dataset with minimal false negatives. Such a high recall is essential for building a reliable detection system that captures all forms of offensive speech, particularly important for safeguarding younger users in online environments.

The F1-score, which combines precision and recall, provided a comprehensive view of the model’s overall performance. Beginning at 0.4814 in the first epoch, the F1-score surged to 0.9887 by epoch 15 and remained consistently high, ranging between 0.97 and 1.00 through to the final epoch. A high and stable F1-score suggests that the model excels not only in one aspect (either precision or recall) but maintains a balanced ability to distinguish between toxic and non-toxic comments. This strongly supports the viability of using ViT-based approaches in image-based toxic language detection systems, particularly on platforms like TikTok.

Discussion

The Vision Transformer (ViT) model used in this study demonstrated excellent performance in detecting offensive comments on TikTok, as evidenced by significant improvements in key evaluation metrics accuracy, precision, recall, and F1-score over 50 training epochs. In the initial stages, the model’s accuracy was relatively low at 63.24% in the first epoch. However, performance improved rapidly, and by epoch 15, the accuracy had surpassed 99%. After this point, the model's performance stabilized, maintaining accuracy between 98.5% and 99.9% until the end of training. This level of stability indicates that the model not only learns quickly but also retains a high level of classification capability throughout the training process. Given the rising popularity of social media platforms like TikTok, especially among younger generations(Krisdanu & Kiranastari Asoka Sumantri, 2023; Wallace, 2024) ,such detection capabilities are crucial.

Precision, as an indicator of how accurately the model predicts the positive class (i.e., offensive comments), also showed a steep increase. From a low value of 0.5190 in the first epoch, the model's precision steadily rose to 0.9898 by epoch 15. This high level of precision suggests that the model rarely produces false positives neutral comments are seldom misclassified as offensive. This is essential for ensuring that moderation systems do not overly filter comments that are actually harmless. In a previous study, Dwitama (2021) emphasized that minimizing false positives is critical in hate speech detection on social media platforms.

Recall also experienced significant improvement. Beginning at 0.5094 in the first epoch, it increased to 0.9877 by epoch 15 and reached a perfect score of 1.000 by epoch 40. A recall of 1.000 indicates the model’s ability to identify nearly all offensive comments in the dataset without missing any (i.e., very low false negatives). With such a high recall, the detection system becomes highly reliable in capturing all forms of offensive speech crucial for creating a safer digital environment, particularly for younger users. Jadmiko & Damariswara, (2022) also noted that children and adolescents are especially vulnerable to the influence of harsh language online, highlighting the need for early detection.

The F1-score, a harmonic mean of precision and recall, provided a comprehensive overview of the model’s performance. Starting from 0.4814 at the beginning of training, it rose sharply to 0.9887 by epoch 15 and remained consistently high in the range of 0.97 to 1.00 until the final epoch. A high and stable F1-score indicates that the model performs well not just in one area (either precision or recall) but maintains a balanced ability to detect both offensive and non-offensive comments. This strengthens the confidence that ViT-based approaches are effective for image-based toxic language detection systems, especially on platforms like TikTok. According to Kurniawan & Mustikasari, (2021), deep learning-based visual processing approaches have proven effective in addressing issues of harmful content, including image-based comments frequently found on TikTok(Alifah Arde Ajeng Hamidah et al., 2023)

Although achieving perfect recall and F1-score values (1.000) may seem impressive, this condition needs to be critically analyzed because training and evaluation were performed on the same dataset without a validation split. This situation increases the risk of overfitting, where the model does not truly learn generalized patterns but merely memorizes the characteristics of the training data (Zhu et al., 2023). This aligns with the findings of Steiner et al.(2022), which emphasize that Vision Transformer (ViT) models are highly dependent on dataset size and prone to overfitting when data is limited and augmentation or regularization strategies are not optimally applied. Furthermore, demonstrated that the absence of a validation set makes overfitting detection difficult, thus additional monitoring or alternative methods such as gradient-based early stopping should be considered. A perfect F1-score on training data can also be misleading because this metric does not necessarily reflect the model’s generalization ability (Chicco & Jurman, 2020). Therefore, employing stricter evaluation strategies, such as using a validation split or applying cross-validation techniques, is crucial to ensure that the reported performance truly represents the model’s generalization capability in real-world scenarios.

4. CONCLUSION

This study successfully implemented the Vision Transformer (ViT) model to detect offensive language on the TikTok social media platform using an image-based approach. Utilizing a dataset of 1,401 classified screenshots of comments, the model demonstrated exceptionally high performance during training over 50 epochs without an explicit validation data split. The achieved metrics include an accuracy of 99.93%, precision of 0.9979, recall of 1.000, and an F1-score of 1.000. These results indicate that ViT is highly effective in recognizing visual patterns associated with offensive language, consistent with previous studies showing that Vision Transformers excel in capturing long-range visual dependencies and can achieve high performance across various image classification tasks (Chhabra & Kumar Vishwakarma, 2024; Liu et al., 2022; Touvron et al., 2021).

The model exhibited stable performance without significant fluctuations during training, indicating that overfitting did not occur despite the absence of a separate validation set. Empirical studies on ViT training have shown that, with proper regularization and augmentation techniques, ViT can achieve high performance even on medium-sized datasets without experiencing performance degradation due to overfitting (Steiner et al., 2021). Additionally, integrated research has revealed that the phenomenon of patch-level overfitting in ViT can be mitigated through dynamic augmentation strategies such as MixUp and PatchErasing, which enhance the model's generalization to visual variations (Zhang et al., 2022). These findings demonstrate that an image-based approach, when combined with well-managed training strategies, can be an effective alternative for detecting offensive language on platforms like TikTok. The model is capable of recognizing offensive content directly from image representations without relying on raw text data.

For further development, it is necessary to evaluate the model’s generalization capability through cross-validation or a more systematic data split, as well as expand the dataset with more complex visual variations. This approach can serve as a foundation for adaptive and effective visual-based automatic content moderation systems across various social media platforms, as also demonstrated in the TikGuard study, which utilized a transformer-based model to detect inappropriate video content for children on TikTok (Balat et al., 2024)

Furthermore, to enhance the robustness of content moderation systems, future research should explore the integration of visual and textual modalities into a unified multimodal detection system. Architectures such as VAuLT have demonstrated success in combining text and images for sentiment analysis tasks on social media, highlighting the effectiveness of this approach in handling complex data (Chochlakis et al., 2022). Similarly, the Semantic Fusion method, which merges visual and textual inputs, has proven to improve the validation of multimedia moderation tools (Wang et al., 2023). Another study, such as METER, illustrates efficient fusion strategies for processing multimodal data in end-to-end transformer-based models (Dou et al., 2022). Therefore, implementing a content moderation system that integrates NLP and computer vision can provide a more comprehensive and context-aware framework, which is highly relevant for platforms like TikTok with visually rich content.

ACKNOWLEDGEMENTS

Thank you to all parties who have provided support and assistance throughout the research process, whether in the form of guidance, motivation, or facilities that enabled this research to be completed successfully. Special thanks are extended to the academic supervisor for their valuable guidance and feedback, as well as to colleagues who contributed to the data collection and labeling processes. It is hoped that the results of this research will provide meaningful contributions to the development of more adaptive content moderation technologies in today's digital era.

REFERENCES

Abdullah, P. M. (2015). Living in the world that is fit for habitation : CCI's ecumenical and religious relationships. In *Aswaja Pressindo*.

Alifah Arde Ajeng Hamidah, Sinta Rosalina, & Slamet Triyadi. (2023). Kajian Sosiolinguistik Ragam Bahasa Gaul di Media Sosial Tiktok pada Masa Pandemi Covid-19 dan Pemanfaatannya Sebagai Kamus Bahasa Gaul. *Jurnal Onoma: Pendidikan, Bahasa, Dan Sastra*, *9*(1), 61–68. https://doi.org/10.30605/onoma.v9i1.2029

Balat, M., Gabr, M., Bakr, H., & Zaky, A. B. (2024). TikGuard: A Deep Learning Transformer-Based Solution for Detecting Unsuitable TikTok Content for Kids. *NILES 2024 - 6th Novel Intelligent and Leading Emerging Sciences Conference, Proceedings*, 337–340. https://doi.org/10.1109/NILES63360.2024.10753192

Bilali, A., Katsiroumpa, A., Koutelekos, I., Dafogianni, C., Gallos, P., Moisoglou, I., & Galanis, P. (2025). Association Between TikTok Use and Anxiety, Depression, and Sleepiness Among Adolescents: A Cross-Sectional Study in Greece. *Pediatric Reports*, *17*(2). https://doi.org/10.3390/pediatric17020034

Chhabra, A., & Kumar Vishwakarma, D. (2024). *MHS-STMA: Multimodal Hate Speech Detection via Scalable Transformer-Based Multilevel Attention Framework*.

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, *21*(1), 1–13. https://doi.org/10.1186/s12864-019-6413-7

Chochlakis, G., Srinivasan, T., Thomason, J., & Narayanan, S. (2022). *VAuLT: Augmenting the Vision-and-Language Transformer for Sentiment Classification on Social Media*. http://arxiv.org/abs/2208.09021

Conte, G., Iorio, G. Di, Esposito, D., Romano, S., Panvino, F., Maggi, S., Altomonte, B., Casini, M. P., Ferrara, M., & Terrinoni, A. (2025). Scrolling through adolescence: a systematic review of the impact of TikTok on adolescent mental health. *European Child and Adolescent Psychiatry*, *34*(5), 1511–1527. https://doi.org/10.1007/s00787-024-02581-w

Contreras Ortiz, A., Santiago, R. R., Hernandez, D. E., & Lopez-Montiel, M. (2025). Multiclass Evaluation of Vision Transformers for Industrial Welding Defect Detection. *Mathematical and Computational Applications*, *30*(2), 1–21. https://doi.org/10.3390/mca30020024

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). an Image Is Worth 16X16 Words: Transformers for Image Recognition At Scale. *ICLR 2021 - 9th International Conference on Learning Representations*.

Dou, Z. Y., Xu, Y., Gan, Z., Wang, J., Wang, S., Wang, L., Zhu, C., Zhang, P., Yuan, L., Peng, N., Liu, Z., & Zeng, M. (2022). An Empirical Study of Training End-to-End Vision-and-Language Transformers. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2022-June*, 18145–18155. https://doi.org/10.1109/CVPR52688.2022.01763

Dwitama, A. P. J. (2021). Deteksi Ujaran Kebencian Pada Twitter Bahasa Indonesia Menggunakan Machine Learning: Reviu Literatur. *Jurnal Sains, Nalar, Dan Aplikasi Teknologi Informasi*, *1*(1). https://doi.org/10.20885/snati.v1i1.5

Islam, K. (2022). *Recent Advances in Vision Transformer: A Survey and Outlook of Recent Work*. *50*, 1–7. http://arxiv.org/abs/2203.01536

Jadmiko, R. S., & Damariswara, R. (2022). Analisis Bahasa Kasar yang Ditirukan Anak Remaja dari Media Sosial Tiktok di Desa Mojoarum Kecamatan Gondang Kabupaten Tulungagung. *Stilistika: Jurnal Pendidikan Bahasa Dan Sastra*, *15*(2), 227. https://doi.org/10.30651/st.v15i2.13162

Jannah, K. A. M., Aiman, U., Hasda, S., Fadilla, Z., Ardiawan, T. M. K. N., & Sari, M. E. (2022). Metodologi Penelitian Kuantitatif Metodologi Penelitian Kuantitatif. In *Metodologi Penelitian Kuantitatif* (Issue May).

Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., & Shah, M. (2022). Transformers in Vision: A Survey. *ACM Computing Surveys*, *54*(10), 1–30. https://doi.org/10.1145/3505244

Krisdanu, C. A., & Kiranastari Asoka Sumantri. (2023). TikTok sebagai Media Pemasaran Digital di Indonesia. *Jurnal Lensa Mutiara Komunikasi*, *7*(2), 24–36. https://doi.org/10.51544/jlmk.v7i2.4173

Kurniawan, A. A., & Mustikasari, M. (2021). Implementasi Deep Learning Menggunakan Metode CNN dan LSTM untuk Menentukan Berita Palsu dalam Bahasa Indonesia. *Jurnal Informatika Universitas Pamulang*, *5*(4), 544. https://doi.org/10.32493/informatika.v5i4.6760

Kurniawan, A. W., & Puspitaningtyas, Z. (2023). Metode Penelitian Kuantitatif (Edisi Revisi). In *Yayasan Kita*

Menulis (Vol. 4, Issue 1).

Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., Wei, F., & Guo, B. (2022). Swin Transformer V2: Scaling Up Capacity and Resolution. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2022-June*, 11999–12009. https://doi.org/10.1109/CVPR52688.2022.01170

Nabila Budihartana, S., & Sudrajat, R. H. (2024). *Pengaruh Konten Tiktok Akun @Imeyhou Terhadap Etika Komunikasi Remaja*. *11*(6), 6921.

Rendragraha, A. D., Bijaksana, M. A., & Romadhony, A. (2021). Pendekatan Metode Transformers untuk Deteksi Bahasa Kasar dalam Komentar Berita Online Indonesia. *E-Proceeding of Engineering*, *8*(2), 3385–3395.

Romindo, R., Pangaribuan, J. J., & Barus, O. P. (2023). Implementasi Algoritma Tf-Idf Dan Support Vector Machine Terhadap Analisis Pendeteksi Komentar Cyberbullying Di Media Sosial Tiktok. *Device*, *13*(1), 124–134. https://doi.org/10.32699/device.v13i1.5260

Steiner, A., Kolesnikov, A., Zhai, X., Wightman, R., Uszkoreit, J., & Beyer, L. (2022). How to train your ViT? Data, Augmentation, and Regularization in Vision Transformers. *Transactions on Machine Learning Research*, *2022-May*(18).

Sujarweni, V., & Wiratna. (2023). *Akuntansi Sektor Publikif: Untuk Mahasiswa Psikologi*.

Sulistiyawati, P., Alzami, F., Prabowo, D. P., Pramunendar, R. A., Megantara, R. A., Purinsyira, N., & Irawan, E. (2022). Prediksi Kata Kasar Berbahasa Indonesia Menggunakan Machine Learning Berbasis Mobile Infrastructure. *Transmisi*, *24*(2), 55–61. https://doi.org/10.14710/transmisi.24.2.55-61

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2021). Training data-efficient image transformers & distillation through attention. *Proceedings of Machine Learning Research*, *139*, 10347–10357.

Wallace, A. R. (2024). "on the Tendency of Varieties To Depart Indefinitely From the Original Type." *Evolution in Victorian Britain: Volume I: Evolution before Darwin*, *1*, 371–379. https://doi.org/10.4324/9781003490548-32

Wang, W., Huang, J., Chen, C., Gu, J., Zhang, J., Wu, W., He, P., & Lyu, M. (2023). Validating Multimedia Content Moderation Software via Semantic Fusion. *ISSTA 2023 - Proceedings of the 32nd ACM SIGSOFT International Symposium on Software Testing and Analysis*, *1*, 576–588. https://doi.org/10.1145/3597926.3598079

Zhang, Q., Xu, Y., Zhang, J., & Tao, D. (2023). ViTAEv2: Vision Transformer Advanced by Exploring Inductive Bias for Image Recognition and Beyond. *International Journal of Computer Vision*, *131*(5), 1141–1162. https://doi.org/10.1007/s11263-022-01739-w

Zhu, H., Chen, B., & Yang, C. (2023). *Understanding Why ViT Trains Badly on Small Datasets: An Intuitive Perspective*. 1–10. http://arxiv.org/abs/2302.03751