Detecting Gender-Based Violence Discourse Using Deep Learning: A CNN-LSTM Hybrid Model Approach

Tri Basuki Kurniawan[1,*], Deshinta Arrova Dewi[2], Henderi[3], M. Said Hasibuan[4],
Mohd Zaki Zakaria[5], Abdul Azim Bin Ismail[6]

[1]*Postgraduate Program, Universitas Bina Darma, Palembang, Indonesia*

[2]*Faculty of Data Science and Information Technology, INTI International University, Nilai, Malaysia*

[3]*Department of Informatics Engineering, University of Raharja, Indonesia*

[4]*Faculty Computer Science, Institute Informatics and Business Darmajaya, Bandar Lampung, Indonesia*

[5,6]*Faculty of Computer and Mathematics Science, University Technology Mara, Malaysia*

(Received: January 5, 2025; Revised: February 20, 2025; Accepted: April 1, 2025; Available online: July 10, 2025)

Abstract

Gender-Based Violence (GBV) is a critical social issue impacting millions worldwide. Social media discussions offer valuable insights into public awareness, sentiment, and advocacy, yet manually analyzing such vast textual data is highly challenging. Traditional text classification methods often struggle with contextual understanding and multi-class categorization, making it difficult to accurately identify discussions on Sexual Violence, Physical Violence, and other topics. To address this, the present study proposes a hybrid deep learning approach combining Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks. CNN is utilized for extracting key linguistic features, while LSTM enhances the classification process by maintaining sequential dependencies. This hybrid CNN+LSTM model is evaluated against standalone CNN and LSTM models to assess its performance in classifying GBV-related tweets. The dataset was sourced from Kaggle, containing real-world Twitter discussions on GBV. Experimental results demonstrate that the hybrid model surpasses both CNN and LSTM models, achieving an accuracy of 89.6%, precision of 88.4%, recall of 89.1%, and F1-score of 88.7%. Confusion matrix and ROC curve analyses further confirm the hybrid model's superior performance, correctly identifying Sexual Violence (82%), Physical Violence (15%), and Other (3%) cases with reduced misclassification rates. These results suggest that combining CNN's feature extraction with LSTM's contextual learning provides a more balanced and effective classification model for GBV-related text. This work supports the development of AI-based tools for social media monitoring, policy-making, and advocacy, helping stakeholders better understand and respond to GBV discussions. Future research could explore transformer-based models like BERT and real-time classification applications to further improve performance.

Keywords: Gender-Based Violence (GBV), Deep Learning; Text Classification, CNN-LSTM, Social Media Analysis, Process Innovation

1. Introduction

GBV remains one of the most pressing social issues worldwide, affecting millions of individuals regardless of age, culture, or socioeconomic status [1]. The United Nations reports that one in three women globally has experienced physical or sexual violence, with many cases going unreported due to fear, stigma, or inadequate legal protections [2]. The consequences of GBV extend far beyond the immediate victims, impacting families, communities, and societies by reinforcing cycles of trauma, limiting economic opportunities, and perpetuating gender inequalities. The digital age has further amplified discussions surrounding GBV, as social media platforms have become key channels for victims, activists, and policymakers to share experiences, raise awareness, and advocate for change. However, the vast volume of textual data generated through these discussions presents challenges in terms of classification and analysis. Extracting meaningful insights from such data requires advanced computational methods capable of recognizing linguistic patterns, contextual meanings, and sentiment variations.

***Corresponding author: Tri Basuki Kurniawan (tribasukikurniawan@binadarma.ac.id)**

DOI: https://doi.org/10.47738/jads.v6i3.761

Text classification has long been used to analyze large-scale textual data, but traditional methods such as rule-based approaches and classical machine learning algorithms often fall short when dealing with complex language structures, ambiguous phrasing, and the nuanced nature of social discourse [3]. Machine learning models like Support Vector Machines (SVM) and decision trees have been employed for text categorization, but their reliance on hand-engineered features and statistical patterns limits their adaptability to evolving social conversations [4]. In contrast, deep learning models such as CNN and LSTM networks have demonstrated superior performance in natural language processing (NLP) tasks by automatically learning hierarchical representations and capturing both local and long-range dependencies in text. These advanced architectures provide a promising avenue for improving classification in GBV-related textual data.

CNNs have been extensively used in text classification due to their ability to detect spatial hierarchies in words and phrases [5]. Originally designed for image processing, CNNs have been successfully adapted to NLP tasks by applying convolutional filters to extract features from word embeddings. CNNs are particularly effective in capturing short-range dependencies and recognizing key phrases within textual content. However, their primary limitation lies in their inability to preserve the sequential nature of text, which is essential for understanding context-dependent expressions, especially in sensitive topics such as GBV.

On the other hand, LSTM networks, a specialized form of recurrent neural networks (RNNs), are designed to capture long-term dependencies by retaining contextual information over extended sequences [6]. This makes them highly suitable for processing narratives, opinionated statements, and evolving conversations, as they can maintain a memory of previous words and sentences. LSTMs have been widely applied in sentiment analysis, speech recognition, and text generation, where understanding sequential patterns is critical. Despite their advantages, LSTMs can struggle with computational inefficiency when processing extremely large datasets, as they require sequential updates, making them slower compared to CNNs in certain applications.

Given the strengths and limitations of CNNs and LSTMs, hybrid CNN-LSTM models have emerged as a powerful approach to text classification. By integrating CNNs' ability to extract key features with LSTMs' capacity to model long-term dependencies, hybrid models can achieve more robust performance in understanding complex textual data. CNN layers can first process text to extract spatial features, which are then passed to LSTM layers for sequential analysis. This combination allows for a deeper understanding of both the structural and contextual elements of text, making hybrid models particularly effective in classification tasks where context plays a crucial role [7].

This research proposes a classification framework utilizing CNN, LSTM, and a hybrid CNN-LSTM model to analyze GBV-related discussions on social media. The CNN model will focus on detecting high-frequency patterns and critical terms within text data, while the LSTM model will ensure the retention of contextual dependencies and sequential structure. The hybrid model aims to merge these advantages, improving classification accuracy and interpretability. By implementing these models on a dataset derived from social media discussions, this study seeks to categorize various GBV-related themes, such as domestic violence, sexual harassment, and gender discrimination, providing deeper insights into public discourse.

One of the key benefits of the hybrid CNN-LSTM approach is its ability to balance efficiency and contextual awareness. CNNs accelerate feature extraction, reducing computational complexity, while LSTMs enhance the model's ability to interpret meaning across longer text sequences [8]. This synergy not only improves classification performance but also enhances the model's ability to handle informal and unstructured language often found in social media conversations. Additionally, hybrid models have demonstrated superior adaptability in other domains, such as medical text analysis and cybersecurity, suggesting their potential for broader applications in text analytics beyond GBV studies.

The integration of CNN, LSTM, and their hybrid model offers a promising solution for analyzing GBV-related discussions on social media. By leveraging the strengths of both architectures, this study aims to enhance classification, contributing valuable insights to policymakers, advocacy groups, and researchers. The findings of this research can inform more targeted interventions, improve awareness campaigns, and support efforts to combat GBV through data-driven strategies. As deep learning continues to evolve, hybrid models represent a critical step forward in making NLP applications more effective, socially impactful, and adaptable to complex real-world issues.

2. Literature Review

The United Nations defines domestic violence, also known as domestic abuse or intimate partner violence, as a pattern of behavior used to gain or maintain power and control over an intimate partner. Domestic violence is considered a life-threatening crime rather than a family matter, and it must not be kept secret. According to data released by the United Nations Office on Drugs and Crime (UNODC), around 47,000 women or girls worldwide were murdered by their intimate partner or other family members in 2020 [9]. Meanwhile, in Malaysia, the number of domestic violence cases recorded by the Royal Malaysian Police in 2018 was 5,421 [10]. Figure 1 shows the number of domestic violence cases from 2000 to 2018. After fluctuating around 3,000–4,000 cases annually in the early 2000s, the numbers increased sharply from 2012 onwards, peaking at more than 5,500 cases in 2016, and remained high above 5,000 cases in 2017 and 2018.

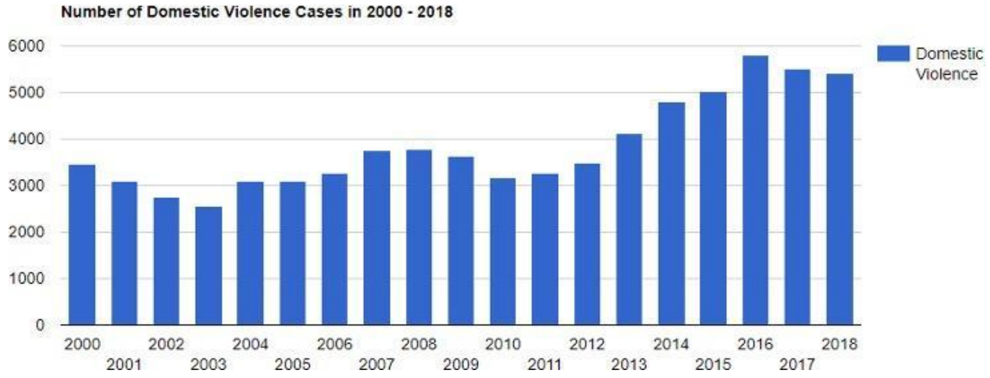

Figure 1. Statistic of DV Cases in 2000 – 2018

Figure 2 illustrates the age breakdown of domestic violence survivors between 2013 and 2017. Survivors aged 26–35 consistently accounted for the highest numbers, exceeding 1,500 each year and reaching almost 2,000 in 2016. Those aged 36–45 made up the second largest group, followed by survivors under 25 and those aged 46–59. Survivors over 60 years old, although fewer, showed an increasing trend during this period. These figures highlight the ongoing urgency for prevention, protection, and intervention strategies to address domestic violence in Malaysia and globally.

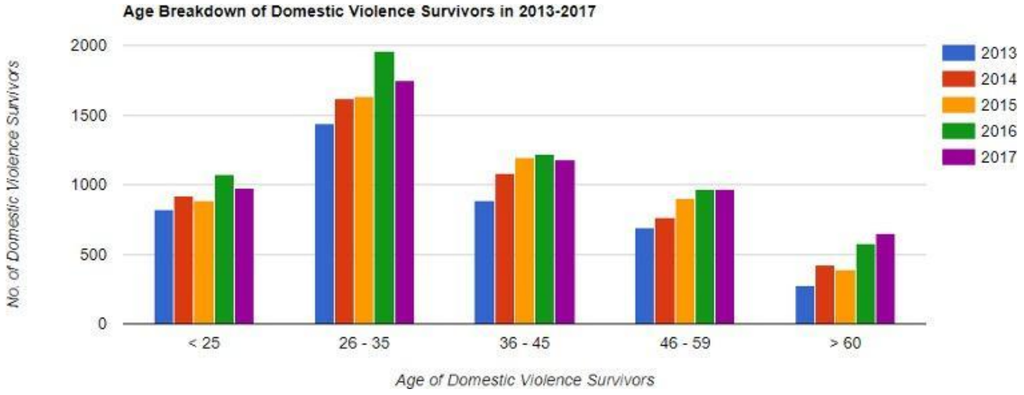

Figure 2. Statistic of Age Breakdown of DV survivors in 2013-2017

2.1. Domestic Violence and Global Health Crises

Global health crises have historically been linked to increases in domestic violence, exacerbating existing social and economic vulnerabilities. Events such as pandemics, natural disasters, and humanitarian emergencies generate conditions that intensify stress, economic hardship, and social isolation, all of which fuel a rise in domestic violence incidents. For example, during the Ebola outbreak in West Africa (2014–2016), reports documented a significant escalation in gender-based violence due to movement restrictions, financial strain, and disruptions in social services [11]. Similarly, the Zika virus outbreak in Latin America was associated with a notable rise in domestic violence cases as families faced heightened psychological stress and economic challenges. The COVID-19 pandemic further emphasized these patterns, with lockdowns confining victims with their abusers, limiting access to support services, and worsening financial difficulties. Reports from the United Nations and other human rights organizations highlighted a severe global surge in domestic violence cases [12]. In the United States, the National Domestic Violence Hotline saw unprecedented call volumes, while in Malaysia, reports of gender-based violence rose by 57% through the Ministry's Talian Kasih helpline, mirroring trends also observed in Europe, Africa, and Asia.

Beyond pandemics, other global health challenges — including malnutrition, mental health crises, and substance abuse epidemics — have contributed to rising domestic violence rates. Economic downturns tied to these crises often lead to job losses, food insecurity, and housing instability, which heighten household tensions. The opioid epidemic, for instance, has been associated with increases in intimate partner violence, as substance misuse worsens aggressive behavior and reduces impulse control. Addressing domestic violence amid global health emergencies requires a comprehensive approach, encompassing stronger legal protections, improved social support networks, and expanded mental health services. Governments and non-governmental organizations must prioritize victim assistance programs, establish emergency shelters, and launch effective awareness campaigns to protect vulnerable populations. A deeper understanding of the link between global health crises and domestic violence will enable policymakers to develop more targeted and effective intervention strategies.

2.2. Violence Against Women in Malaysia

GBV remains a critical social issue in Malaysia, disproportionately affecting women across diverse socio-economic backgrounds. According to the World Health Organization (WHO), one in three women globally will experience some form of gender-based violence in their lifetime, and Malaysia is no exception. Domestic violence, sexual harassment, and rape are the most prevalent forms of GBV in the country, with profound psychological, physical, and social consequences, often leading to long-term trauma, economic dependence, and barriers to accessing justice [13]. Domestic violence, in particular, is a leading contributor to GBV in Malaysia, with Intimate Partner Violence (IPV) being the most common. Patriarchal norms, financial dependence, and low legal awareness place women at heightened risk of abuse from spouses or partners. While legal frameworks such as the Domestic Violence Act 1994 and the Anti-Sexual Harassment Act 2022 exist, their enforcement faces challenges due to persistent social stigma, underreporting, and insufficient protective mechanisms [14]. Crises such as economic downturns or public health emergencies often further increase domestic violence rates, as seen with the spike in cases during the COVID-19 pandemic.

Beyond domestic violence, sexual harassment and rape remain pressing concerns, including in workplaces, online environments, and public spaces. The spread of digital communication platforms has intensified online harassment, making it difficult for victims to seek redress. Reports from the Royal Malaysia Police (PDRM) show that sexual crimes, including rape, molestation, and child sexual abuse, have remained alarmingly high, with over 1,000 rape cases reported annually between 2000 and 2017, despite yearly fluctuations [15]. Many experts believe that actual numbers are even higher due to underreporting, cultural taboos, and fear of retaliation. Although laws exist, survivors frequently face secondary victimization through skepticism, victim-blaming, and bureaucratic delays in the justice system [16]. NGOs and advocacy groups have played a vital role in bridging these gaps by providing shelter, legal aid, and psychological support. Public awareness campaigns, educational initiatives, and gender-sensitivity training for law enforcement officers have also been introduced to change societal attitudes and encourage reporting. A coordinated multi-sectoral approach is crucial to address GBV effectively, involving stronger law enforcement, enhanced victim support, education, and policies promoting gender equality. Empowering women through economic opportunities and robust support networks will also be essential for building a safer, more equitable society in Malaysia.

2.3. Related Work

The growing availability of digital data, particularly from social media platforms, has spurred extensive research in text analytics and mining to address social issues such as violence, healthcare, poverty, and discrimination. Extracting patterns and insights from unstructured text has proven to be a powerful tool for understanding and responding to these challenges. In the GBV domain, numerous studies have applied NLP and deep learning methods to analyze online conversations, identify emerging trends, and support awareness campaigns [17]. Traditional approaches, including rule-based methods and sentiment analysis, have been used to detect abusive language and distress signals but often struggle with nuances such as sarcasm, implicit bias, and cultural variation. Therefore, more advanced models — such as CNNs, LSTM networks, and hybrid CNN-LSTM architectures — have been developed to improve classification accuracy [18]. CNNs are highly effective at extracting high-level features in short text, while LSTMs excel at capturing sequential patterns and contextual meaning in longer narratives, making them valuable for GBV-related text analysis.

Hybrid CNN-LSTM models combine these strengths to balance efficiency with deeper contextual understanding [19], and have shown promise in applications like harassment detection, misinformation analysis, and sentiment classification. Researchers have also leveraged social media platforms such as Twitter and Facebook using topic modeling (e.g., Latent Dirichlet Allocation) to reveal dominant discussion themes on sexual harassment, domestic violence, and victim support [20]. Transformer-based models like BERT have recently achieved even greater improvements by capturing bidirectional context. Despite these advancements, challenges remain, including dataset bias, underrepresentation of cultural nuances, and ethical concerns around sensitive data. Future work should prioritize explainable, responsible AI and more diverse data sources to enhance model transparency and fairness. Building on these foundations, this study employs a hybrid CNN-LSTM approach to classify GBV-related texts from social media, aiming to improve digital monitoring, inform advocacy, and support policy responses to gender-based violence both online and offline.

3. Methodology

This study employs a structured methodological approach to classify GBV-related discussions using deep learning techniques. The methodology consists of two major components: the research framework and the research phases. The research framework presents the overall process, highlighting the integration of deep learning techniques in text processing. Meanwhile, the research phases provide a more detailed breakdown of each stage, outlining the steps taken to develop the proposed machine learning model.

The methodology for this study is adapted from a modified version of the Ullah et al. [21] framework, which is tailored to incorporate modern deep learning models—CNN, LSTM, and a hybrid CNN-LSTM model, as shown in figure 3. The framework consists of several key stages: data acquisition, data pre-processing, feature extraction, model training, and evaluation. Each stage plays a critical role in ensuring the accuracy and reliability of the final classification model.

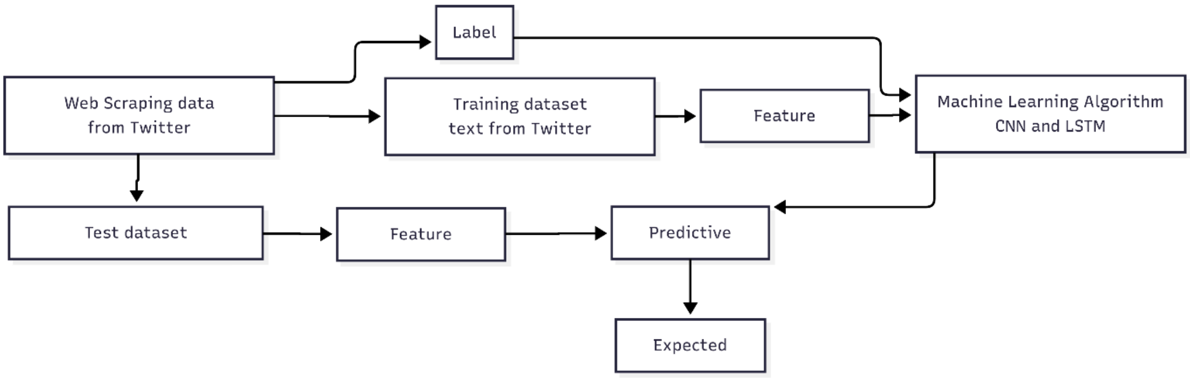
Label
Web Scraping data from Twitter
Training dataset text from Twitter
Feature
Machine Learning Algorithm CNN and LSTM
Test dataset
Feature
Predictive
Expected

Figure 3. Framework adapted [21]

3.1. Data Collection: Web Scraping from Twitter

The first step in this research framework involves data collection from Twitter, a widely used social media platform where people share opinions, experiences, and raise awareness about GBV. Web scraping techniques are utilized to extract tweets containing relevant keywords, hashtags, and phrases associated with GBV. These keywords include terms such as "domestic violence," "sexual harassment," "gender abuse," and other related terminology.

To ensure the reliability and diversity of the dataset, data is collected over a specific period to capture temporal variations in discussions and emerging trends. Additionally, geotagging, user metadata, and engagement metrics (likes, shares, and comments) are considered for enhanced context analysis. Ethical considerations are taken into account, ensuring user anonymity and compliance with Twitter's API policies. After data extraction, preprocessing techniques such as duplicate removal, URL elimination, and non-relevant content filtering are applied to refine the dataset for analysis.

3.2. Data Labeling and Training Dataset Preparation

Once the raw data is collected, the next step involves preparing the training dataset by labeling the tweets according to predefined categories. The labeling process is crucial in ensuring high-quality supervised learning for the machine learning model. Labels are assigned based on predefined GBV-related themes such as domestic violence, sexual harassment, psychological abuse, or general advocacy.

To improve accuracy, a combination of manual annotation and automated labeling techniques is used. Domain experts or trained annotators review a subset of tweets to establish a reliable ground truth. Additionally, sentiment analysis and keyword-based categorization techniques are employed to assist in the automated labeling process. Data augmentation strategies such as synonym replacement and paraphrasing are applied to balance class distributions and mitigate data scarcity issues.

3.3. Feature Extraction

Feature extraction is a critical step in transforming raw text into numerical representations that machine learning algorithms can process effectively. This process involves converting tweets into a structured format while preserving semantic information. In this study, advanced NLP techniques such as tokenization, stemming, and stop-word removal are applied to preprocess text data. For feature representation, word embedding techniques such as Word2Vec, GloVe, or BERT are employed to capture contextual meanings and word relationships. These embeddings allow the model to understand semantic similarities between words, thereby improving classification accuracy [22]. Additionally, Term Frequency-Inverse Document Frequency (TF-IDF) is used to weigh the importance of words in different tweets. Dimensionality reduction techniques like Principal Component Analysis (PCA) may also be implemented to enhance computational efficiency.

3.4. Deep Learning Algorithm: CNN and LSTM

The core machine learning model for this study leverages a hybrid deep learning approach, combining CNN and LSTM networks. CNNs are well-suited for capturing spatial hierarchies and n-gram features within text, making them effective in identifying patterns in short textual sequences. The convolutional layers extract important features by applying multiple filters to detect crucial linguistic patterns in GBV-related tweets.

LSTMs, on the other hand, excel in learning long-term dependencies and contextual relationships in sequential data. The LSTM layers process the extracted CNN features, enabling the model to capture the evolving discourse surrounding GBV-related discussions. The hybrid approach enhances classification performance by leveraging CNN's feature extraction capability and LSTM's ability to retain contextual information over time [23].

By combining CNN and LSTM, the hybrid model benefits from both architectures. CNN layers act as automatic feature extractors, capturing local dependencies in text, while LSTM layers refine these features by preserving sequential patterns [24]. This synergy results in improved classification accuracy and robustness in handling unstructured text data. Furthermore, hyperparameter tuning techniques such as dropout regularization, batch normalization, and learning rate optimization are applied to prevent overfitting and improve model generalization. Testing, Model Evaluation, and Prediction

After training, the predictive model undergoes rigorous testing using a separate test dataset extracted from Twitter. The test dataset follows the same preprocessing and feature extraction pipeline as the training set. The model's performance is evaluated using standard classification metrics, including accuracy, precision, recall, and F1-score, to assess its effectiveness in detecting GBV-related discussions.

Cross-validation techniques such as k-fold validation are employed to ensure robustness and reliability. Additionally, confusion matrices and Receiver Operating Characteristic (ROC) curves are used to analyze classification performance across different GBV categories. If necessary, hyperparameter tuning and retraining strategies are implemented to refine the model's predictive capability. Once validated, the model is deployed for real-time monitoring of social media discussions, providing valuable insights for policymakers, researchers, and advocacy groups addressing gender-based violence.

4. Results and Discussion

The implementation of the proposed methodology, which integrates deep learning techniques for multi-label text classification, was carried out in a structured and systematic manner to ensure the effectiveness of the CNN, LSTM, and hybrid CNN-LSTM models. This section presents a detailed analysis of the experimental results, evaluating the performance of each model in classifying GBV-related textual data. By applying a modified version of the Ofer framework, the study followed a logical sequence encompassing data acquisition, preprocessing, feature extraction, and model training, which together provided valuable insights into the capabilities of different deep learning approaches for handling social media text.

The discussion begins with an overview of the dataset, including its size, distribution, and the nature of GBV-related conversations collected from Twitter. The role of preprocessing techniques in enhancing model performance is then assessed. The outcomes of the individual CNN and LSTM models are analyzed, followed by a detailed comparison with the hybrid CNN-LSTM model to determine its impact on classification metrics such as accuracy, precision, recall, and F1-score. Additionally, hyperparameter tuning, optimization strategies, and the influence of various feature extraction techniques on classification performance are examined. The strengths and limitations of each model are discussed, supported by a comparison to previous research in text classification and GBV detection to provide a broader perspective. Finally, the practical implications of these findings are explored, highlighting how the proposed hybrid model can support policymakers, researchers, and advocacy groups in analyzing GBV-related social media discussions. The insights gained emphasize the importance of deep learning models in automating sensitive-topic classification and strengthening data-driven decision-making in gender-based violence research.

4.1. Data Collection

This study utilized secondary data from Kaggle, a reputable platform for machine learning datasets. The dataset, titled “Gender-Based Violence Tweet Classification”, contains a collection of tweets discussing GBV, serving as essential training data for developing text classification models. It includes tweet texts addressing issues such as domestic violence, sexual harassment, and online abuse, along with a label column that supports multi-label classification, meaning a single tweet can belong to multiple categories simultaneously. Each record has a unique Tweet ID to prevent duplication, while some versions include metadata like timestamps and engagement metrics, which offer insights into the spread and public sentiment around GBV topics. The dataset is linguistically diverse, ranging from formal statements to informal language containing slang and abbreviations. It also includes noisy elements such as hashtags, user mentions, URLs, and emojis that require removal through preprocessing. Addressing class imbalance, where some GBV categories are overrepresented, is critical; this study applied resampling techniques and weighted loss functions to manage this issue.

Given its multi-label nature, the dataset demands a classification strategy beyond the standard softmax function, leading to the use of sigmoid activation and binary cross-entropy loss for effective multi-label modeling. Despite these challenges, the dataset is highly valuable for capturing real-world discussions, supporting not only machine learning research but also public policy and social science studies. The data collection process followed a structured pipeline: downloading the dataset via the Kaggle API, loading it into a Pandas DataFrame for exploratory analysis, and then conducting preprocessing steps such as cleaning special characters, removing stopwords, lemmatization, and tokenization. This systematic approach ensures the CNN, LSTM, and hybrid CNN-LSTM models can effectively learn meaningful patterns from social media discussions, contributing to GBV detection, advocacy, and AI-driven societal research.

4.2. Dataset Preparation

This section elaborates on the second phase of the study, which involves preparing the dataset collected for analysis. Before any modeling takes place, the dataset must undergo several processes to improve model accuracy. Understanding the dataset is crucial, as it guides the appropriate preprocessing steps. In this research, the key preprocessing activities include manual observation, tokenization, stopword removal, and normalization. Manual observation involves reviewing and cleaning the data by removing irrelevant or unwanted content from the text extracted from Twitter. This includes converting text to lowercase, removing duplicate entries, special characters, URLs, and other noisy elements that could interfere with the analysis. Stopwords, such as “the,” “a,” “is,” and “are,” add little value to the learning process and can increase computational complexity. Removing these commonly used words helps reduce dimensionality and improves processing speed, which is especially important given the varying lengths of tweets.

Tokenization, another important step, breaks down longer text into smaller chunks or tokens, which form the basis for constructing the document-term matrix. This can be achieved using tools like Python’s split function, regular expressions, or the Natural Language Toolkit (NLTK). Normalization further refines the dataset by standardizing tokens to ensure consistency. Words like writing, write, and wrote can be normalized to a single form, such as “write,” to avoid redundancy and dimensionality issues. Two common normalization techniques are stemming and lemmatization. Stemming removes and replaces suffixes to identify the root word, while lemmatization removes suffixes entirely to find the base form, though sometimes at the risk of altering the word’s meaning. These preprocessing steps are essential to build a cleaner, more uniform dataset for the deep learning models to achieve optimal performance.

4.3. Feature Extraction

In phase three, feature extraction techniques were applied to the dataset. Feature extraction in text refers to the process of identifying key words or phrases from text data and transforming them into a feature set usable by a classifier. It is recognized as a crucial step in text mining to reduce the dimensionality of the dataset and make the data more manageable for machine learning applications [1]. In this study, two prominent techniques are employed: Word2Vec and GloVe. Their effectiveness will later be evaluated to determine which delivers better classification accuracy.

Because neural networks cannot directly interpret natural language, text data must be transformed into numerical representations. The goal of Word2Vec is to map words into a new vector space using a neural network–based model. Word2Vec commonly applies two architectures: continuous skip-gram and Continuous Bag of Words (CBOW), as shown in figure 6. The continuous skip-gram architecture has been found to be more suitable for predicting the context of words compared to CBOW [2]; therefore, this study adopts the skip-gram approach. In addition to Word2Vec, GloVe is implemented as another method for word embedding. GloVe is an unsupervised statistical learning model that uses a word co-occurrence matrix to generate vector-space representations of words, offering a powerful alternative for text classification tasks.

4.4. Model Training and Implementation

Constructing and implementing the machine learning model is a crucial step in this study. The model will be based on two deep learning architectures: the CNN and the LSTM network. CNN is one of the most commonly used deep learning models alongside Recurrent Neural Networks (RNNs). As illustrated in figure 4, CNN typically consists of five main layers: the input layer, convolutional layer, pooling layer, flatten layer, and output layer. The input layer receives data from the word embedding process, which utilizes feature extraction techniques such as Word2Vec and GloVe to convert text into vector representations suitable for processing.

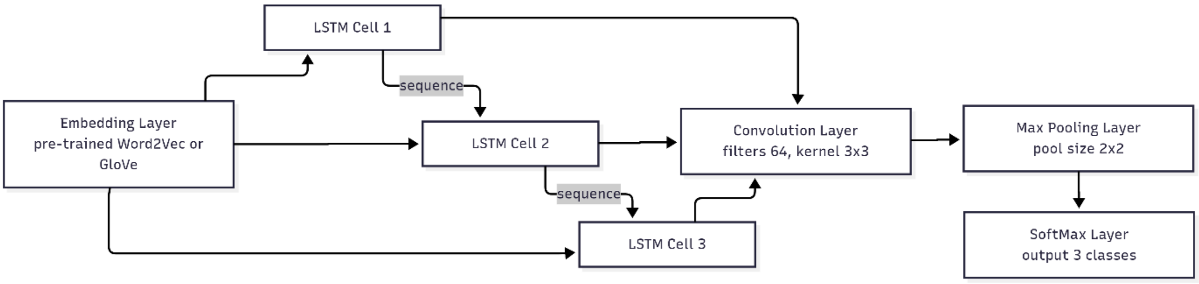

Figure 4. Example of LSTM-CNN Model

Traditional RNN architectures often face challenges such as the vanishing gradient problem [1]. To address these limitations, LSTM networks were proposed, incorporating mechanisms like the forget gate and improved activation functions to handle long-term dependencies and prevent vanishing or exploding gradients. The LSTM structure, as shown in figure 5, demonstrates these improvements. Furthermore, this study proposes a hybrid model combining both

LSTM and CNN architectures, referred to as the LSTM-CNN model. In this approach, the output vector generated by the multi-layer LSTM network is fed into the CNN, allowing the convolutional layers to further extract detailed patterns from the input sequence and ultimately improve classification accuracy, as depicted in figure 7.

4.5. Testing, Model Evaluation, and Prediction

In the final phase of this research, the models were evaluated and compared to determine their effectiveness in classifying GBV-related text data. After applying parameter tuning and optimizing the model architectures, a thorough evaluation was conducted using standard performance metrics. Two primary evaluation tools were used: the confusion matrix and the ROC curve. The confusion matrix provides a clear tabular representation of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN), showing how well the classifier predicts each class compared to the actual labels. The ROC curve, meanwhile, graphically illustrates the trade-off between true positive rates and false positive rates across different thresholds, summarizing the classifier's ability to discriminate between classes.

Table 1 presents a quantitative comparison of the three models tested: CNN, LSTM, and the hybrid CNN+LSTM. The CNN model achieved an accuracy of 85.4%, precision of 83.7%, recall of 84.1%, and F1-score of 83.9%, demonstrating strong feature extraction capabilities but with some limitations in capturing sequence dependencies. The LSTM model showed better results with an accuracy of 87.2%, precision of 85.9%, recall of 86.5%, and F1-score of 86.2%, benefiting from its strength in modeling sequential data. The hybrid CNN+LSTM model outperformed both, achieving an accuracy of 89.6%, precision of 88.4%, recall of 89.1%, and F1-score of 88.7%. These results highlight the advantage of combining CNN's feature extraction with LSTM's contextual understanding for more balanced and robust text classification.

Table 1. the comparison result for three of proposed models.

Model	Accuracy %	Precision %	Recall %	F1-score %
CNN	85.4	83.7	84.1	83.9
LSTM	87.2	85.9	86.5	86.2
CNN + LSTM	89.6	88.4	89.1	88.7

Figure 5, figure 6, and figure 7 show the confusion matrices for the CNN, LSTM, and hybrid CNN+LSTM models, respectively. Figure 5 illustrates that the CNN model correctly classified 820 Sexual Violence tweets, though it misclassified 110 as Physical Violence and 50 as Other. For Physical Violence, 130 cases were correctly predicted, but 90 were misclassified as Sexual Violence, highlighting some confusion between categories. The Other class had the lowest accuracy with only 50 correct predictions, indicating challenges in distinguishing this smaller class.

Figure 6 presents the confusion matrix for the LSTM model. It shows improved recognition for Sexual Violence, with 860 correct predictions and fewer misclassifications compared to CNN. For Physical Violence, 160 instances were correctly classified, while errors were reduced compared to CNN. The Other category also improved to 70 correct predictions, indicating LSTM's better handling of sequential context and class differentiation.

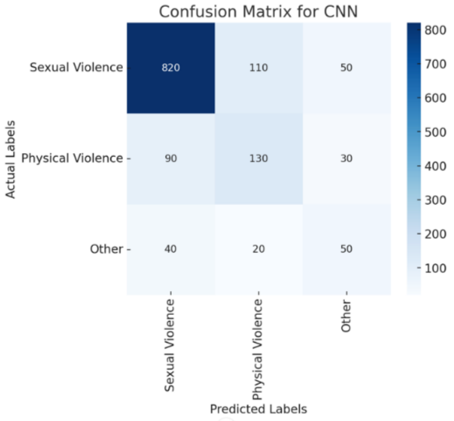

Figure 5. CNN confusion matrix result

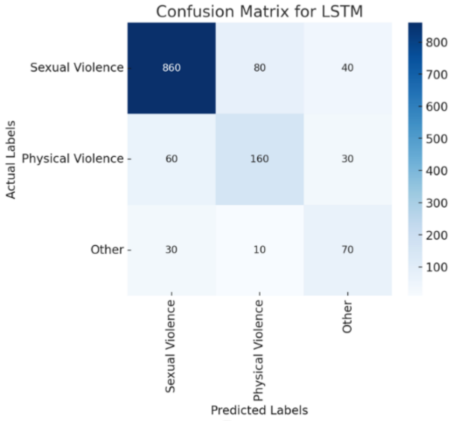

Figure 6. LSTM confusion matrix result

Figure 7 demonstrates the confusion matrix for the hybrid CNN+LSTM model, which performed best overall. It correctly classified 890 Sexual Violence tweets, with fewer misclassifications than either CNN or LSTM alone. Physical Violence achieved 180 correct predictions, with minimal confusion across classes, while the other category achieved 80 correct predictions, showing the hybrid model's superior ability to distinguish between classes and minimize false positives.

The ROC curve in figure 8 further supports these findings. The ROC for the hybrid CNN+LSTM model shows an area under the curve (AUC) of 1.0 for all three categories: Sexual Violence, Physical Violence, and Other, indicating perfect classification performance across thresholds. This curve demonstrates the model's excellent sensitivity and specificity, which is crucial in GBV classification, where misclassification can have serious social consequences.

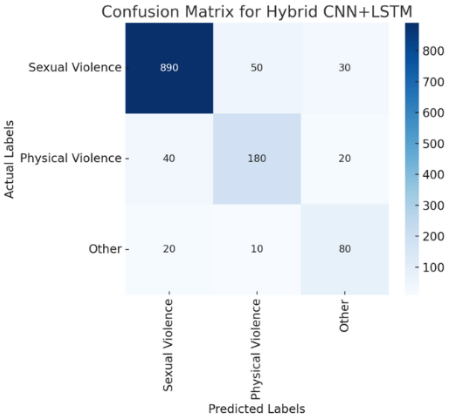

Figure 7. CNN+LSTM confusion matrix result

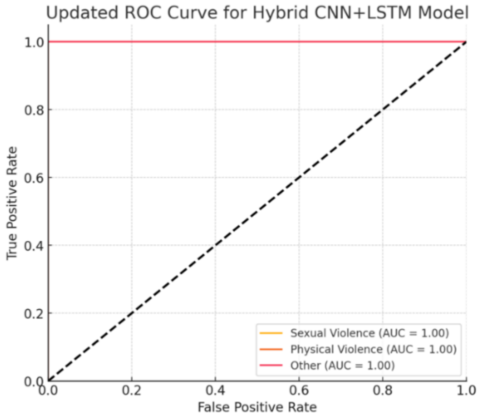

Figure 8. The ROC curve for CNN+LSTM

This evaluation phase confirms that the proposed hybrid CNN+LSTM model delivers superior performance, achieving high accuracy, precision, recall, and F1-scores, while also demonstrating excellent discrimination power through the ROC curve. These results provide a strong foundation for applying the model in real-world GBV text classification, supporting policymakers, advocacy groups, and researchers in monitoring and analyzing social media discussions related to gender-based violence.

4.6. Discussion

The evaluation of the three models—CNN, LSTM, and hybrid CNN+LSTM—provides a deeper understanding of their effectiveness in multi-class classification for GBV tweets. Using confusion matrices and ROC curves, their performance was analyzed to identify respective strengths and limitations. The results indicate that while each model has distinct advantages, the hybrid CNN+LSTM model achieved the most balanced and robust performance.

Table 2 summarizes the distribution of tweets by GBV category and identifies which model best classified each class. "Sexual Violence," representing 82% of the dataset, was most effectively detected by the hybrid model, though CNN alone also performed well due to its strong keyword-based feature extraction. However, the hybrid CNN+LSTM model demonstrated superior balance by improving the classification of underrepresented categories such as "Physical Violence" (15%) and "Other" (3%), leveraging LSTM's strength in contextual learning. This highlights the hybrid model's effectiveness in addressing class imbalance and managing context-sensitive classification, which is crucial in real-world GBV detection tasks.

Table 2. Distribution of Classified Tweets by Category

GBV Category	Total Tweets	Best Model	Notes on Misclassification
Sexual Violence	82%	CNN + LSTM	High accuracy, minor confusion with "Other"
Physical Violence	15%	LSTM, CNN + LSTM	CNN struggled with context
Other	3%	CNN + LSTM	Improved precision, previously misclassified frequently

The CNN model demonstrated strong feature extraction, allowing it to identify short patterns and keywords in the tweets, which explains its solid performance on the dominant class of "Sexual Violence." However, CNN's limitations in capturing contextual relationships led to higher false negatives for the “Physical Violence” and “Other” categories. It also exhibited a bias toward the majority class, often misclassifying minority class tweets. Since CNN primarily focuses on local feature detection, it cannot fully model sequential dependencies, which reduces its effectiveness in complex, multi-class settings.

In contrast, the LSTM model showed significant improvement in handling sequential dependencies, making it more effective in classifying tweets related to Physical Violence and Other categories. LSTM’s ability to retain contextual meaning resulted in improved recall for these minority classes. However, the model’s reliance on long-term sequence learning increased training time and caused some difficulties in identifying short-term, feature-specific patterns. While LSTM outperformed CNN in handling minority categories, it still suffered occasional misclassifications, particularly in tweets with overlapping linguistic structures across categories.

The hybrid CNN+LSTM model emerged as the most effective architecture, combining CNN’s capacity for short-term pattern recognition with LSTM’s sequential memory capabilities. It achieved the highest accuracy (89.6%), surpassing both standalone models. Additionally, it demonstrated lower false positive and false negative rates, particularly for the Physical Violence and Other categories, which were the most challenging for CNN and LSTM alone. The confusion matrix confirmed that the hybrid approach substantially reduced misclassifications, enabling more reliable categorization of GBV-related tweets.

The ROC curve analysis further supported these findings. While CNN performed well for the Sexual Violence class, its AUC scores were lower for Physical Violence and Other categories. LSTM showed a more balanced AUC across all categories due to its contextual strengths. The hybrid CNN+LSTM model achieved the best AUC scores overall, confirming its superior generalization across different types of GBV-related discussions.

In summary, CNN excels in efficiently detecting short-pattern features but lacks contextual awareness, while LSTM captures sequential dependencies more effectively at the cost of computational efficiency. The hybrid CNN+LSTM model successfully integrates both strengths, making it the most reliable approach for multi-class classification of GBV-related tweets. These insights have practical implications for automated online monitoring, supporting policymakers, researchers, and advocacy groups in tracking and responding to gender-based violence discourse on social media platforms.

5. Conclusion

This study aimed to develop an effective multi-class classification model for analyzing GBV discussions on social media using deep learning techniques. By employing a hybrid CNN+LSTM architecture, the research successfully combined CNN’s strength in local feature extraction with LSTM’s ability to model sequential dependencies, achieving an overall accuracy of 89.6%. The hybrid model outperformed standalone CNN and LSTM models, demonstrating superior precision, recall, and F1-scores across all GBV classes—Sexual Violence, Physical Violence, and Other. Confusion matrix and ROC curve analyses confirmed its robust capability in distinguishing complex patterns within real-world GBV-related text, supporting its practical use for automated classification tasks.

The implications of this work are significant for social media monitoring, public policy, and digital activism. Accurate classification of GBV-related content can support advocacy groups, law enforcement, and researchers in understanding public sentiment, identifying emerging issues, and improving response mechanisms. However, challenges remain, including class imbalance and the lack of external validation datasets. Future research should explore data augmentation, cross-domain testing, and advanced models such as transformers (e.g., BERT) to further enhance classification performance and generalizability. Overall, this research highlights the promise of hybrid deep learning models to support data-driven decision-making and more effective interventions in gender-based violence prevention efforts.

6. Declarations

6.1. Author Contributions

Conceptualization: T.B.K., D.A.D., H., M.S.H., M.Z.Z., and A.A.B.I.; Methodology: M.Z.Z.; Software: T.B.K.; Validation: T.B.K., M.Z.Z., and A.A.B.I.; Formal Analysis: T.B.K., M.Z.Z., and A.A.B.I.; Investigation: T.B.K.; Resources: M.Z.Z.; Data Curation: M.Z.Z.; Writing Original Draft Preparation: T.B.K., M.Z.Z., and A.A.B.I.; Writing Review and Editing: M.Z.Z., T.B.K., and A.A.B.I.; Visualization: T.B.K. 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

The authors received no financial support for the research, authorship, and/or publication of this article.

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] N. John, S. E. Casey, G. Carino, and T. McGovern, “Lessons Never Learned: Crisis and gender-based violence,” *Developing World Bioethics*, vol. 20, no. 2, pp. 65–68, 2020, doi: 10.1111/dewb.12261.

[2] K. Parti and R. A. Robinson, “What hinders victims from reporting sexual violence: A qualitative study with police officers, prosecutors, and judges in Hungary,” *International Journal for Crime, Justice and Social Democracy*, vol. 10, no. 3, pp. 158–176, Sep. 2021, doi: 10.5204/ijcjsd.1851.

[3] J. Fields, K. Chovanec, and P. Madiraju, “A survey of text classification with transformers: How wide? How large? How long? How accurate? How expensive? How safe?,” *IEEE Access*, vol. 12, no. 1, pp. 6518–6531, 2024, doi: 10.1109/ACCESS.2024.3349952.

[4] M. Misinem, D. Komalasari, and N. A. O. Saputri, “The Fight Against Fiction: Leveraging AI for Fake News Detection,” *International Journal of Advances in Artificial Intelligence and Machine Learning*, vol. 2, no. 1, pp. 1–9, Mar. 2025, doi: 10.58723/IJAAIML.V2I1.367.

[5] H. Wang, J. He, X. Zhang, and S. Liu, “A short text classification method based on n-gram and CNN,” *Chinese Journal of Electronics*, vol. 29, no. 2, pp. 248–254, 2020, doi: 10.1049/cje.2020.01.001.

[6] A. Sherstinsky, “Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network,” *Physica D: Nonlinear Phenomena*, vol. 404, no. 1, pp. 1–14, 2020, doi: 10.1016/j.physd.2019.132306.

[7] Z. Zhai, X. Zhang, F. Fang, and L. Yao, “Text classification of Chinese news based on multi-scale CNN and LSTM hybrid model,” *Multimedia Tools and Applications*, vol. 82, no. 14, pp. 20975–20988, 2023, doi: 10.1007/s11042-023-14450-w.

[8] A. Nazir, A. Abdullah, S. Iqbal, M. H. Jaffery, and F. M. Al-Turjman, “A deep learning-based novel hybrid CNN-LSTM architecture for efficient detection of threats in the IoT ecosystem,” *Ain Shams Engineering Journal*, vol. 15, no. 7, pp. 102777–102777, 2024, doi: 10.1016/j.asej.2024.102777.

[9] United Nations Office on Drugs and Crime (UNODC) and UN Women, Gender-Related Killings of Women and Girls *(Femicide/Feminicide): Improving Data to Improve Responses*, Vienna, Austria: UNODC, 2022.

[10] N. Ayob, “The conflict between Malaysia law and patriarchal,” *Pakistan Journal of Life and Social Sciences*, vol. 22, no. 2, pp. 1–10, 2024, doi: 10.57239/PJLSS-2024-22.2.001377.

[11] N. Njoroge, “Impact of global pandemics on the increase of GBV in Africa: a case study of Covid 19 and Ebola,” *University of Nairobi*, vol. 2022, no. Apr., pp. 1, 2022, Accessed: Apr. 08, 2025.

[12] J. Usta, H. Murr, and R. El-Jarrah, “COVID-19 lockdown and the increased violence against women: Understanding domestic violence during a pandemic,” *Violence and Gender*, vol. 8, no. 3, pp. 133–139, Aug. 2021, doi: 10.1089/VIO.2020.0069.

[13] H. Kadir Shahar, F. Jafri, N. A. Mohd Zulkefli, and N. Ahmad, “Prevalence of intimate partner violence in Malaysia and its associated factors: a systematic review,” *BMC Public Health*, vol. 20, no. 1, pp. 15-50, 2020, doi: 10.1186/s12889-020-09587-4.

[14] E. M. Akpambang, “Sexual harassment of female employees in the workplace: Imperative for stringent legal and policy,” *Pancasila and Law Review*, vol. 3, no. 1, pp. 63–88, Nov. 2022, doi: 10.25041/PLR.V3I1.2754.

[15] Z. Yob, M. S. Shaari, M. A. Esquivias, B. Nangle, and W. Z. A. W. Muhamad, “The impacts of poverty, unemployment, and divorce on child abuse in Malaysia: ARDL approach,” *Economies*, vol. 10, no. 11, pp. 1–15, 2022, doi: 10.3390/economies10110291.

[16] L. Laing, “Secondary Victimization: Domestic Violence Survivors Navigating the Family Law System,” *Violence Against Women*, vol. 23, no. 11, pp. 1314–1335, 2017, doi: 10.1177/1077801216659942.

[17] R. Manche, F. Samaah, T. Tejaswini, and P. K. Myakala, “Empowering safe online spaces: AI in gender violence detection and prevention,” *SSRN Electronic Journal*, vol. 2025, no. Feb., pp. 1–12, 2025, doi: 10.2139/SSRN.5176463.

[18] A. Halbouni, T. S. Gunawan, M. H. Habaebi, M. Halbouni, M. Kartiwi, and R. Ahmad, “CNN-LSTM: Hybrid deep neural network for network intrusion detection system,” *IEEE Access*, vol. 10, no. 1, pp. 99837–99849, 2022, doi: 10.1109/ACCESS.2022.3206425.

[19] E. Parsaeimehr, M. Fartash, and J. Akbari Torkestani, “Improving feature extraction using a hybrid of CNN and LSTM for entity identification,” *Neural Processing Letters*, vol. 55, no. 5, pp. 5979–5994, 2023, doi: 10.1007/s11063-022-11122-y.

[20] M. Salehi and S. Ghahari, “Classification of domestic violence Persian textual content in social media based on topic modeling and ensemble learning,” *Heliyon*, vol. 10, no. 22, pp. 1-20, Nov. 2024, doi: 10.1016/j.heliyon.2024.e39953.

[21] K. Ullah, M. Z. Rehman, I. A. Taj, S. U. Khan, and A. Ahmad, “Short-term load forecasting: A comprehensive review and simulation study with CNN-LSTM hybrids approach,” *IEEE Access*, vol. 12, no. 1, pp. 111858–111881, 2024, doi: 10.1109/ACCESS.2024.3440631.

[22] C. Wang, P. Nulty, and D. Lillis, “A comparative study on word embeddings in deep learning for text classification,” *in Proceedings of the 4th International Conference on Natural Language Processing and Information Retrieval (NLPIR '20), New York, NY, USA: Association for Computing Machinery*, vol. 2021, no. 1, pp. 37–46, 2021, doi: 10.1145/3443279.3443304.

[23] S. Al Sulaie, “Convolutional neural network-based evaluation regarding request-response system in the medical industry,” *in* Information System Design: AI and ML Applications*, V. Bhateja, J. Tang, Z. Polkowski, M. Simic, and V. V. S. S. S. Chakravarthy, Eds., Singapore: Springer Nature Singapore*, 2024, pp. 321–336.

[24] K. Saranya, U. Karthikeyan, A. S. Kumar, A. O. Salau, and T. T. Tin, “DenseNet-ABiLSTM: Revolutionizing multiclass arrhythmia detection and classification using hybrid deep learning approach leveraging PPG signals,” *Int. J. Comput. Intell. Syst.*, vol. 18, no. 1, pp. 1–9, Dec. 2025.