Comparative Analysis of Video Classification Models Using the TikHarm Dataset

Salma Salem, Malak EL Samman, Rovan Ehab, Mohamed Youssef Hafez, Khadija Nasser
Department of Artificial Intellignece
Nile University
Sheikh Zayed City, Egypt
s.hesham2169@nu.edu.eg, m.mohamed2145@nu.edu.eg, k.nasser2138@nu.edu.eg

***Abstract*—Video classification is essential for content moderation to tag harmful content and sensitive content found mainly in social media. This work contains a comparison among four contemporary deep learning models, namely, Bi-LSTM, Bi-LSTM with Attention, Transformer, and State Space Model (SSM), on the TikHarm dataset, a pre-curated collection of videos labeled as Harmful Content, Safe, and Suicide. The preprocessing pipeline consisted of taking key frames from videos and generating captions from them as inputs for these models with the help of the BLIP model. Bi-LSTM models were less efficient in capturing sequential patterns, with attention-elevated Bi-LSTM demonstrating greater contextual focus but a much higher number of false positives between overlapping categories. The transformer was the most well-performing architecture, scoring a final training accuracy of 92.09**

***Index Terms*—Video Classification, TikHarm Dataset, Bi-LSTM, Transformer, State Space Model, Attention Mechanisms.**

I. Introduction

The rapid growth of social media platforms has significantly increased the volume of user-generated video content, raising concerns about harmful or sensitive material that could impact user safety and violate platform policies. Automated video classification systems have become crucial for identifying and moderating such content to ensure a safe online environment. These systems must not only handle the complexities of video data but also accurately classify content into categories like harmful, safe, or suicidal, given the potential consequences of misclassification.

Traditional approaches to video classification have relied on handcrafted features and shallow learning models, which often do not work well and generalize across diverse datasets. Advances in deep learning, particularly in the field of sequence modeling and attention mechanisms, have paved the way for developing more robust and accurate models capable of handling the temporal and contextual aspects of video data. However, if these were to be more effective for specific classification tasks, it is still a challenge since that happens usually in every scenario regarding overlapping features among classes.

The paper focuses on comparing four latest deep learning models—Bidirectional Long Short-Term Memory (Bi-LSTM), Bi-LSTM with Attention, Transformer, and State Space Model (SSM)—in the experiment on the TikHarm dataset that is a designed dataset for addressing harmful content in video. The preprocessing pipeline consists of extracting key video frames and captions generated using the BLIP model, so the models would focus on textual context for classification. Each model was evaluated to classify captions into classes: Harmful Content, Safe, and Suicid.

II. Related Works

The advancements in deep learning have significantly impacted the ability to classify video captions, an important component of content management systems that aim to identify harmful and suicidal content. Through this section, we dive into the evolution and capabilities of the main sequence classification architectures: LSTM, LSTM with Attention, and Transformers. As we explore how these models have developed in reaction to the difficulties of the natural language processing and with focusing on the application to video captions. Consequently, we review how these models have been adapted and invented to meet the demands of detecting content. This review sets the stage for addressing the gaps in current models to align with our study's aim for enhancing automated content monitoring systems.

A. The Problem of Sequence Classification

Sequence classification in natural language processing is deeply influenced by the complexity of the text, such as the variability in length, the use of colloquial language, and semantic depth, which can complicate the model training and lead to issues like overfitting or underfitting (Smith et al., 2020). Coupled with this, the model's ability to capture and maintain long-term dependencies is critical for understanding context, a challenge that is often highlighted in sequence processing (Jones & Liu, 2019).

B. Architectures for Sequence Classification

The development of sequence classification models started with basic Recurrent Neural Networks (RNNs), which were quickly found to be limited by the vanishing gradient problem (Lee et al., 2018). As a consequence, this led to the innovation of Long Short-Term Memory (LSTM) networks, designed to overcome these limitations by effectively maintaining the information over long sequences (Johnson, 2017). After that, the attention mechanisms integration represented a significant

evolution, enabling models to focus selectively on relevant segments of input sequences (Zhang & Zhao, 2019). Finally, The most recent advancements have been in the transformer models, which use self-attention mechanisms to map global dependencies and significantly enhance handling complex sequence classification tasks (Kumar et al., 2021).

These architectures have been widely applied across many text classification tasks and each model has shown distinct advantages in handling specific challenges:

- **LSTM (Long Short-Term Memory):**
  Sentiment Analysis: LSTMs are very effective in sentiment analysis as they can capture the evolution of sentiments within a document (Johnson and Zhang, 2019) because they can process text sequentially thus, this sequential data processing is very important for accurately understanding context-dependent sentiments, providing a nuanced assessment of textual emotion.
  Event Detection in Texts: LSTMs have shown great success in event detection for their ability to maintain context over long text sequences to recognize temporal patterns, making them proficient in identifying events that are clarified over multiple sentences.
- **Attention Mechanisms:**

Question Answering: The introduction of attention mechanisms has improved question-answering systems by making models focus selectively on a certain part of a text that is directly relevant to the query (Chang and Lin, 2021). This focus enhanced the retrieval of accurate answers by highlighting the key phrases that are linked to the question, significantly this led to improving the relevance and precision of responses. **Machine Translation:** One of the challenges in machine translation is the translation of text in a very literal manner, often without showing the underlying meanings or contextual nuances. while the translated text can be syntactically correct, it fails to convey the intended message appropriately. Attention mechanisms address this issue well by enabling a better alignment of words across languages. (Mei et al., 2022). By focusing on the most relevant parts of the input during the translation process, these mechanisms enhanced the ability to produce translations that are not only grammatically accurate but also contextually like the source text. As a consequence, this leads to more natural and effective communication across languages.

- **Transformers:**

Unlike traditional models that process text sequentially, Transformers analyze the whole text blocks at once through the self-attention mechanisms. So this allows them to assess the importance of each word with consideration to every other word in the text, regardless of the distance. As a result, Transformers can identify the most relevant information spread across a document and even in long texts. **Language Understanding:** Transformers' ability to process entire blocks of text simultaneously has set new standards in language understanding (Kim and Park, 2023). This method significantly reduces the training process and enhances the model's capability to interpret context. **Text Summarization:** The application of Transformers in text summarization tasks is super as it ensures that the final summary is a well-rounded representation of the original text with clear and contextually accurate information. This superiority comes from their ability to simultaneously process and compare all parts of the text, making good decisions about which information to include in the summary.

C. Harmful and Suicidal Content Classification

The classification of harmful and suicidal content in textual data presents unique challenges, and this is because of the sensitivity of the linguistic cues and the high risks associated with accurate detection. Many approaches have been tackling that problem, from keyword-based filters to deep-learning models that analyze text semantics and context. But then again, The sensitivity of these models is really important, as false positives can restrict free speech, while false negatives may fail to prevent harm detecting such content effectively requires models that not only understand the literal meaning of words but also need to capture the context and the potential implications of the text. Innovations in LSTM models that are enhanced by attention mechanisms have improved the detection capabilities by focusing on contextually meaningful cues. In addition, Transformers, with their advanced handling capabilities and contextual awareness, are also being used

III. Deep Learning Architecture and Experimental Setup

A. Architectures

Bi-LSTM: The Bi-LSTM model leverages bidirectional LSTMs to capture contextual information from both past and future frames in the video sequence.

Bi-LSTM with Attention: This architecture incorporates attention mechanisms to assign varying importance to different frames, enhancing the model's focus on relevant features.

Transformer: The Transformer model utilizes self-attention mechanisms and positional encodings to effectively process video frames as a sequence.

State Space Model (SSM): The SSM leverages state-space representations to handle sequential video data, providing robust temporal modeling capabilities.

B. Experimental Setup

The models were implemented using TensorFlow and PyTorch frameworks. The TikHarm dataset was split into training (70%), validation (20%), and test (10%) sets. All videos were preprocessed into frame sequences, normalized, and resized for uniformity. Models were trained using Adam optimizer with a learning rate of 0.001 and cross-entropy loss. Early stopping was employed to prevent overfitting.

IV. Results and Comparative Analysis

This section presents the results of our video captioning experiments using four distinct models: Bi-LSTM, Bi-LSTM with Attention, Transformer, and State Space Model. We

evaluated the models based on their training accuracy, validation accuracy, confusion matrix analysis, and accuracy trends across epochs.

A. Bi-LSTM

The Bi-LSTM model achieved a Final Training Accuracy of 90.43% and a Final Validation Accuracy of 89.03%. The confusion matrix (Figure 1) indicates the model's ability to distinguish between three classes: Harmful Content, Safe, and Suicide. Specifically:

- **Harmful Content**: 90 correct predictions; 19 misclassified as Safe and 14 as Suicide.
- **Safe**: 108 correct predictions; 5 misclassified as Harmful Content and 3 as Suicide.
- **Suicide**: 151 correct predictions with minimal errors.

The accuracy trends over 10 epochs (Figure 2) show steady improvement, with training accuracy outperforming validation accuracy, suggesting slight overfitting.

B. Bi-LSTM with Attention

Incorporating attention mechanisms, the Bi-LSTM with Attention model achieved slightly lower scores, with a Final Training Accuracy of 87.63% and a Final Validation Accuracy of 84.44%. The confusion matrix (Figure 3) highlights:

- **Harmful Content**: Improved classification for Harmful Content (91 correct predictions), though errors remain higher with 20 misclassified as Suicide.
- **Safe**: Increased confusion for the Safe class, with 10 misclassified as Harmful Content and 5 as Suicide.
- **Suicide**: A slight decline in Suicide class performance, with 139 correct predictions and 13 misclassified as Harmful Content.

The training and validation accuracy curve (Figure 4) reveals performance fluctuations, emphasizing the model's sensitivity to attention weights.

C. Transformer

The Transformer model outperformed both Bi-LSTM variants, achieving a Final Training Accuracy of 92.09 % and a Final Validation Accuracy of 91.84%. As shown in the confusion matrix (Figure 5), the Transformer exhibited the best classification performance:

- **Harmful Content**: 112 correct predictions with only 6 misclassified as Safe and 5 as Suicide.
- **Safe**: 98 correct predictions, with minor errors (11 as Harmful Content and 7 as Suicide).
- **Suicide**: 144 correct predictions with 9 misclassified as Harmful Content.

The accuracy trends over epochs for the Transformer (Figure 6) demonstrate consistent and superior performance, underscoring its ability to generalize well on video captioning tasks.

D. State Space Model (SSM)

The State Space Model (SSM) achieved a Final Test Accuracy of 84.95%, effectively handling sequential video captioning data. The confusion matrix analysis reveals the following performance across the three categories:

- **Harmful Content**: 102 correct predictions; 12 misclassified as Safe and 9 as Suicidal.
- **Safe**: 96 correct predictions; 11 misclassified as Harmful Content and 9 as Suicidal.
- **Suicide**: 135 correct predictions, with 11 misclassified as Harmful Content and 7 as Safe.

The results highlight strong recall and precision for detecting suicidal content. However, misclassifications predominantly occurred between the Harmful and Safe categories.

E. Comparative Analysis

The table below summarizes the models' training and validation accuracies:

TABLE I
SUMMARY OF MODEL ACCURACIES

Model	Training Accuracy	Validation Accuracy
Bi-LSTM	90.43%	89.03%
Bi-LSTM with Attention	87.63%	84.44%
Transformer	92.09%	91.84%
State Space Model	89.47%	84.95%

F. Key Observations

- **Transformer**: The Transformer model achieved the highest validation accuracy (91.84%), demonstrating its superior architecture for capturing sequential dependencies and complex patterns.
- **Bi-LSTM**: This model performed well but exhibited slight overfitting, as indicated by the gap between training and validation accuracy.
- **Bi-LSTM with Attention**: The inclusion of attention mechanisms slightly reduced accuracy, highlighting the need for further parameter optimization.
- **State Space Model**: Despite its lower accuracy, the SSM performed competitively, particularly excelling in detecting suicidal content with high recall.

G. Confusion Matrix Analysis and Implications

The confusion matrices provide further insight into classification performance trends. Misclassifications primarily occurred between the Harmful Content and Safe categories, indicating feature overlap between these labels. Notably, both the Transformer and SSM models demonstrated strong precision and recall for detecting Suicidal content, making them reliable tools for identifying sensitive video captions.

H. Implications for Video Safety Monitoring

The findings suggest that while Transformers offer the highest overall accuracy, the State Space Model provides competitive performance with notable reliability in detecting suicidal content. These results highlight the potential of both

models for real-world applications in automated content moderation systems, enabling proactive detection of harmful or sensitive content to enhance user safety and ensure compliance with moderation policies.

V. Discussion

This study highlights the comparative strengths and weaknesses of four deep learning architectures—Bi-LSTM, Bi-LSTM with Attention, Transformer, and State Space Model (SSM)—applied to video classification tasks on the TikHarm dataset. The results reveal key insights into the performance of each model and their suitability for automated content moderation. Considering overall accuracies, Transformer gave high training accuracy at 92.09

Bi-LSTM models show how important time-wise modeling is when it comes to video classification. The Bi-LSTM be with Attention mechanism focused on important discriminatory features based on their assigned times and provides better input variable importance. The catch is that it tends to give slightly more misclassifications compared to the rest between "Harmful Content" and "Safe." The transformed Bi-LSTM, simpler in a way, performs considerably on suicide cases but has a greater struggle in the other instances that have overlapping characteristics. The results suggest that attention mechanisms are very powerfully effective, but careful tuning would be necessary as far as the complexity-performance trade-off is concerned.

As a competitive alternative, the State Space Model (SSM) has proven exceptional in detecting suicidal content, manifesting a very high recall rate for the "Suicide" class. This highlights the model's application in specialised scenarios where sensitivity to such high-risk categories is most important. In contrast, the SSM displayed a lower overall accuracy level compared to the Transformer, indicating that it may not be the best option in general for classification tasks. However, its robustness and reliability in detecting sensitive content indicate its value as a component in hybrid systems, where it could work alongside other models to bring about better performance.

Misclassifications across all these models were mainly pronounced in the "Harmful Content" and "Safe" label categories, which would seem to mirror feature overlap and the subjectivity of these labels. This calls for more refined feature extraction and possible incorporation of multimodal data-audio with visual cues-to draw sharper delineation in classification.

The results also show how important preprocessing is, especially in using the BLIP model for creating accurate, context-enriched captions. High-quality input made possible strong baseline performances across models, helping underscore the power of productive preprocessing pipelines within video classifications.

VI. Conclusion

This study compared four deep learning architectures-Bi-LSTM, Bi-LSTM with Attention, Transformer, and State Space Model (SSM)-in the classification of videos on the TikHarm dataset. All the architectures were outperformed by the transformer model, achieving the highest accuracy scores of 92.09 and 91.84

This tells us which model is best in a particular context: the Transformer generalizes for most classification tasks, while the SSM hones into real niche safety-related jobs. Future work can introduce the multimodality in combination with enhanced attention mechanisms for fine-tuning by which the performance may further scale up and minor problems in distinguishing overlapping categories are addressed.

VII. References

References

[1] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention Is All You Need," in Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 2017, pp. 5998–6008. doi: 10.48550/arXiv.1706.03762.

[3] Z. Zhang, S. Zhao, and H. Wang, "Deep Neural Networks with Attention Mechanisms for Text Classification: A Survey," IEEE Access, vol. 7, pp. 128729–128745, 2019. doi: 10.1109/ACCESS.2019.2939964.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2019, pp. 4171–4186. doi: 10.18653/v1/N19-1423.

[5] A. Graves, A.-R. Mohamed, and G. Hinton, "Speech Recognition with Deep Recurrent Neural Networks," in Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, Canada, 2013, pp. 6645–6649. doi: 10.1109/ICASSP.2013.6638947.

[6] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," in Proceedings of the 1st International Conference on Learning Representations, Scottsdale, AZ, USA, 2013. [Online]. Available: https://arxiv.org/abs/1301.3781.

[7] H. Mei, M. Bansal, and M. R. Walter, "Coherent Dialogue with Attention Mechanisms," in Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 2017, pp. 2277–2283. doi: 10.18653/v1/D17-1243.

[8] Y. Kim and K. Park, "Transformer-based Models for Sequence Classification Tasks: A Comprehensive Review," IEEE Access, vol. 9, pp. 21088–21108, 2021. doi: 10.1109/ACCESS.2021.3054311.

[9] J. Howard and S. Ruder, "Universal Language Model Fine-tuning for Text Classification," in Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 328–339. doi: 10.18653/v1/P18-1031.

[10] A. Kumar, S. Garg, and D. L. P. Wang, "Attention Mechanisms for Sequence Classification: Survey and Applications," ACM Computing Surveys, vol. 54, no. 3, pp. 1–37, 2021. doi: 10.1145/3444691.