AIS411-5

Comparative Analysis of Video Classification Models Using the TikHarm Dataset

Abstract

This research investigates the performance of deep learning models for video classification tasks using the TikHarm dataset, aiming to enhance automated content moderation systems and ensure safer online environments.

Context and Purpose: The study focuses on evaluating state-of-the-art models, including Bi-LSTM, Bi-LSTM with Attention, Transformer, and State Space Models, to address the increasing need for robust systems capable of handling sensitive video content categorization.

Multimodal Preprocessing Pipeline: The research employs an advanced preprocessing framework, using the BLIP model to extract key video frames and generate text captions. This ensures accurate and context-rich inputs for classification tasks.

Model Comparative Analysis: The study conducts a comprehensive evaluation of four machine learning architectures, emphasizing their ability to classify video content into three distinct categories: Harmful, Safe, and Suicidal. The Transformer model achieved the highest accuracy, while the State Space Model excelled in identifying high-risk content.

Temporal and Feature Overlap Challenges: Temporal misclassifications and overlapping features between "Harmful" and "Safe" categories were identified as key challenges. This analysis highlights the need for refined feature extraction techniques to improve model precision.

Practical Applications: The findings support the development of real-time content moderation tools, offering practical solutions for platforms to balance user safety with freedom of expression.

Future Directions: Incorporating multimodal data, such as audio and visual cues, along with enhanced attention mechanisms, is proposed to further improve model performance and address classification ambiguities.

Introduction

In an era of digital ubiquity, accurately categorizing video content is essential for maintaining safe and appropriate online environments. This research investigates various deep learning models using the TikHarm dataset, focusing on the classification of videos into categories such as Harmful, Safe, and Suicidal. Through this study, conducted within the context of video content on social platforms, we utilize advanced modeling techniques to enhance content moderation tools, contributing to digital safety and compliance.

Objectives:

•**Model Evaluation and Comparison:** The primary objective is to assess the performance of different deep learning architectures—Bi-LSTM, Bi-LSTM with Attention, Transformer, and State Space Model—in accurately classifying video content. This analysis will help determine which model provides the most reliable and efficient results for real-time content moderation.

•**Enhancement of Content Moderation:** By identifying the strengths and weaknesses of each model, this research aims to refine video classification techniques, thus improving the accuracy and responsiveness of content moderation systems.

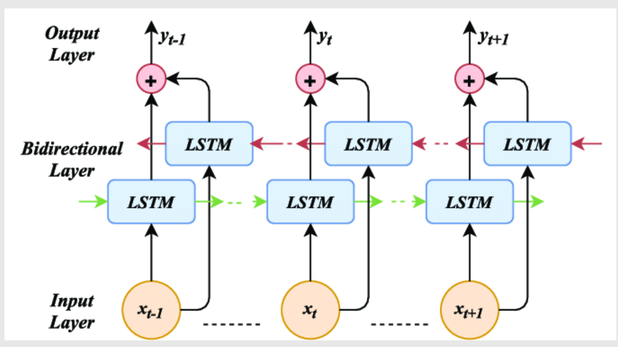
•**Reduction of Misclassification:** A key objective is to minimize the occurrence of misclassifications, particularly between overlapping categories such as "Harmful" and "Safe," enhancing the precision of automated moderation tools.

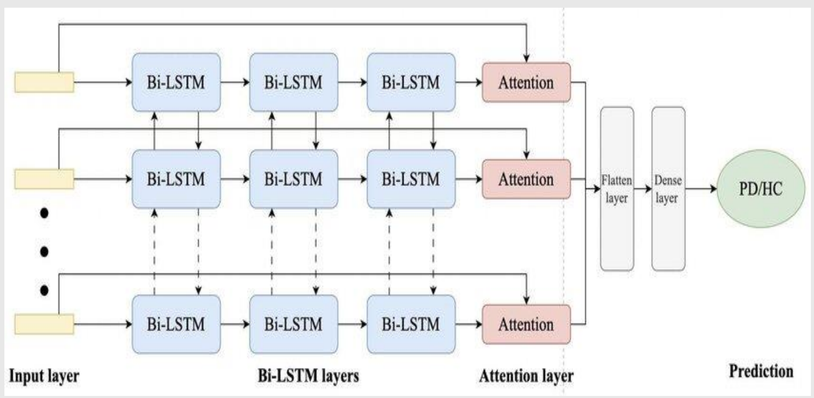
Contributions:

•**Advancing Content Moderation Technologies:** This

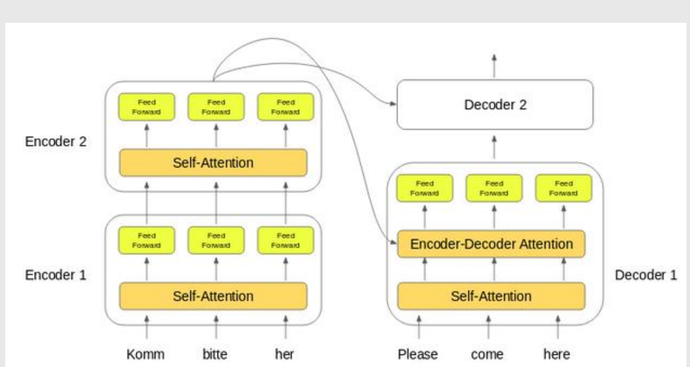
Model Implementation:

•**Bi-LSTM and Bi-LSTM with Attention:** These models process sequences of text extracted from video frames, capturing temporal dependencies. The attention mechanism in the latter model focuses on significant parts of the text to enhance classification accuracy.

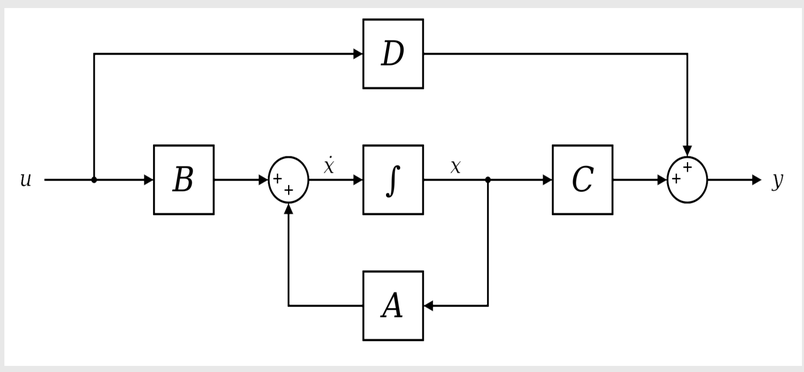




•**Transformer:** This model utilizes self-attention mechanisms to process all words (or frame captions) simultaneously, providing a global understanding of the context, which is crucial for accurately categorizing complex video content.



•**State Space Model (SSM):** SSMs handle video data by modeling the sequence of frames as a dynamic system, which helps in understanding the temporal evolution of video content.



Training Procedure:

•**Hyperparameters:** We employ an Adam optimizer with a learning rate of 0.001, adjusting parameters such as batch size and number of epochs based on preliminary runs to optimize performance.

•**Validation Strategy:** The dataset is split into training (70%), validation (20%), and test (10%) sets. Models are evaluated during

•**Attention Mechanism Impact:** The Bi-LSTM with Attention highlighted its utility in focusing on temporally significant features, which, however, led to a higher rate of misclassifications in overlapping categories. This suggests that while attention mechanisms enhance model sensitivity, they require careful tuning to balance complexity and performance.

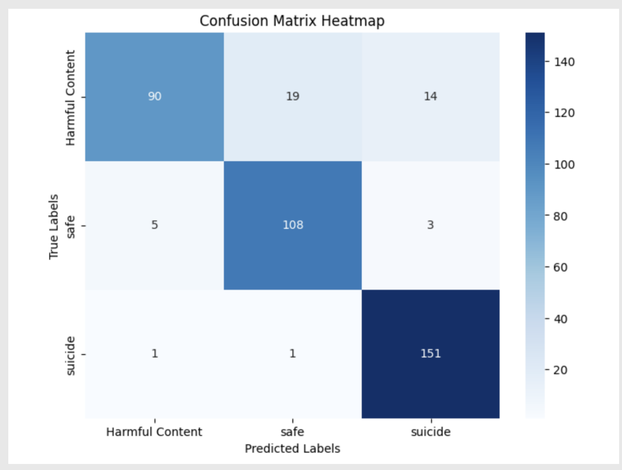
Practical Implications and Recommendations:

•The robustness of the Transformer and the specialized effectiveness of the SSM in certain contexts suggest their potential integration into hybrid systems for improved content moderation.

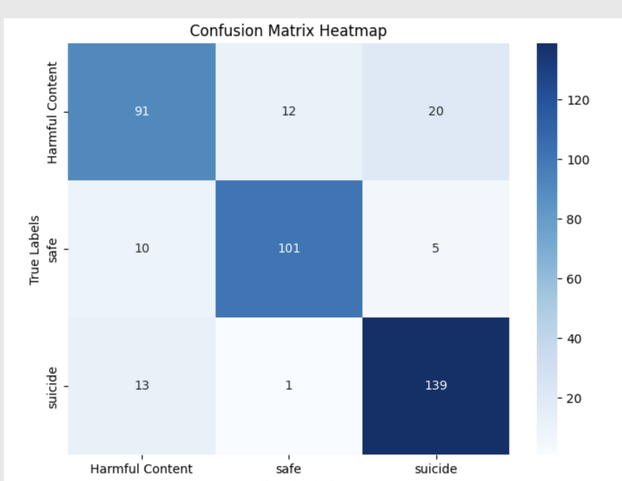
•The need for refined feature extraction and the incorporation of multimodal data (audio and visual cues) are recommended to enhance classification accuracy and reduce misclassifications

TABLE I
SUMMARY OF MODEL ACCURACIES

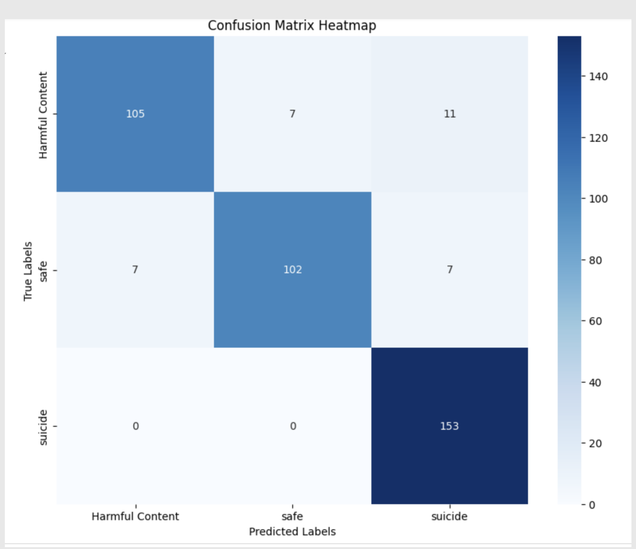
Model	Training Accuracy	Validation Accuracy
Bi-LSTM	90.43%	89.03%
Bi-LSTM with Attention	87.63%	84.44%
Transformer	92.09%	91.84%
State Space Model	89.47%	84.95%



Bi-LSTM



Bi-LSTM with Attention



Transformer