Video Sentiment Analysis for Child Safety

Yee Sen TAN[1], Nicole Anne TEO Huiying[1] Ezekiel En Zhe GHE[1], Jolie Zhi Yi FONG[2], Zhaoxia WANG[1*]
School of Computing and Information Systems, Singapore Management University, Singapore[1]
Lee Kong Chian School of Business, Singapore Management University, Singapore[2]
{yeesen.tan.2020, ezekiel.ghe.2020}@scis.smu.edu.sg, nicolet.2023@engd.smu.edu.sg, jolie.fong.2020@business.smu.edu.sg,
zxwang@smu.edu.sg*

***Abstract*—The proliferation of online video content underscores the critical need for effective sentiment analysis, particularly in safeguarding children from potentially harmful material. This research addresses this concern by presenting a multimodal analysis method for assessing video sentiment, categorizing it as either positive (child-friendly) or negative (potentially harmful). This method leverages three key components: text analysis, facial expression analysis, and audio analysis, including music mood analysis, resulting in a comprehensive sentiment assessment. Our evaluation results validate the effectiveness of this approach, making significant contributions to the field of video sentiment analysis and bolstering child safety measures. This research serves as a valuable resource for those seeking to employ sentiment analysis to protect children from harmful content within the dynamic landscape of video content. Furthermore, our work offers insights into the current state of the art, highlighting the recent advancements, possible improvements, and future directions in video sentiment analysis.**

***Index Terms*—video sentiment analysis, text analysis, facial expression analysis, audio analysis, child safety**

I. Introduction

In today's world, access to the Internet has been greatly enhanced by devices such as smartphones and TVs, which have become highly integrated into our lives. This digital immersion extends even to the youngest members of our society, with children interacting with these devices at increasingly tender ages. Remarkably, most children now own smartphones, exposing them to social media video platforms from an early age. These videos, optimized for mobile consumption, feature fluid transitions enabled by sophisticated recommendation algorithms. These algorithms continuously refine content in real-time based on user interactions, video attributes, and various factors, creating a tailored viewing experience. While recommendation algorithms enhance user experience, concerns arise over negative consequences. Research indicates that social media recommender systems can create filter bubbles, exposing users, especially younger ones, to extreme and disturbing content [1]. This repeated exposure, particularly among young children, raises significant concerns, with potential severe consequences for both physical and mental well-being, including the promotion of self-harm [2]. Recognizing these potential harms, this paper emphasizes the need for astute filtration systems, especially for the substantial audience of young children.

The challenge is intricate, revolving around the identification and filtration of distressing content, recognizing that negativity can permeate through videos, which included the text (e.g., subtitle) in the videos, image (e.g., facial expressions) and audio (including sound and verbal dialogue in videos). The core of this research lies in 'Video Sentiment Analysis', aiming to comprehensively assess video contents. Video sentiment analysis, evolving from text sentiment analysis, adapts to the rise of video content, incorporating text, visual, and audio for a comprehensive assessment [3].

Existing works have demonstrated that the integration and combination of various information sources for video sentiment analysis outperform analyses based solely on text, visual, or audio inputs [4]. Moreover, the multimodal approach has shown considerable promise in sentiment analysis [5].

Our journey takes us deep into the multimodal methodologies employed to extract sentiment from video content, navigating the intricate landscape of textual analysis, facial expression analysis, and audio analysis. Our commitment is to safeguard the digital experiences of children through video sentiment analysis. The main contributions of this paper can be summarized as follows:

- This research proposes a new multimodal analysis method that combines text, facial expressions, and audio (including music mood analysis) for a comprehensive sentiment assessment in online videos.
- The proposed method specifically categorizes video sentiment as positive (child-friendly) or negative (potentially harmful), making contributions to child safety measures in the context of online content.
- Through evaluation results, this research empirically validates the effectiveness of the multimodal approach, making substantial contributions to the field of video sentiment analysis, including insights into recent advancements, potential improvements, and future research directions.
- The research serves as a practical resource for researchers and practitioners, offering valuable tools and methodologies for employing sentiment analysis to protect children from harmful content in the dynamic landscape of online videos.

[*] Corresponding Author: Zhaoxia WANG (e-mail: zxwang@smu.edu.sg)

2375-9259/23/$31.00 ©2023 IEEE
DOI 10.1109/ICDMW60847.2023.00106

II. Literature Review

Sentiment analysis is a natural language processing (NLP) technique that leverages computational methods to determine the polarity or emotional tone expressed in a piece of text [6], [7]. Different AI techniques have been leveraged to improve both accuracy and interpretability of sentiment analysis algorithms, including symbolic AI [8], [9], subsymbolic AI [10], [11], and neurosymbolic AI [12], [13].

Typical applications of sentiment analysis include social network analysis [14], [15], finance [16], and healthcare [17]. Besides traditional algorithms focusing on text [18], multimodal sentiment analysis [19] has also attracted increasing attention recently, driven by the surge in video content consumption and the imperative of ensuring a safe online experience, particularly for children. This section provides an overview of key developments, methodologies, and challenges in video sentiment analysis.

A. Evolution of Video Sentiment Analysis

Video sentiment analysis has its origins in the broader field of sentiment analysis, which initially centered on textual data. However, as video content gained prominence, especially with the proliferation of new user generated videos, video analysis methods must be updated and improved [20]. Researchers sought ways to extract emotions and sentiments embedded within videos, ushering in a multi- modal approach that encompasses visual, auditory, and textual cues to assess emotional content comprehensively.

A multi-head attention-based fusion network was proposed to address the challenge of fusing textual, visual, and acoustic signals for sentiment analysis, and the experimental results demonstrate its good performance over existing methods across multiple modalities [21].

This analysis is complicated by the fact that information is captured in different modalities including text, visual information from the image frames, and audio [22]. In the realm of user-driven short video content, comprehending the interplay between these features is challenging, as users may juxtapose a joyful audio track with melancholic imagery. Nevertheless, incorporating the features of multiple modalities is consistently more accurate than the best unimodal counterparts [23].

B. Multimodal Sentiment Analysis

Multimodal techniques involve integrating individual modalities into a unified model, enabling the extraction of sentiments from diverse sources [24]. Multimodal data offers clear advantages over sole reliance on textual content for analysis [25]. Recent research, such as KnowleNet, also emphasizes multimodal approaches, as it consistently outperforming unimodal models in sarcasm detection [26]. Other existing efforts, like MetaPro, offering potential applications in natural language processing tasks, though it remains an ongoing work [27]. In the realm of multimodal research, the convergence of text and video data, known as text-video retrieval, is crucial.

Traditional methods using global contrastive loss have been criticized for neglecting local alignment and word-level supervision signals. To address these limitations, 'Align and Tell' introduces tailored transformer decoders and local contrastive learning, enhancing retrieval accuracy in multimodal works [28]. In another study, researchers forecasted shifts in audience emotions by leveraging both visual and audio cues. They employed a model architecture for feature extraction and a bidirectional Gated Recurrent Unit (GRU) to capture temporal contextual information effectively. This multimodal approach secured a commendable second-place ranking on the EEV Challenge leaderboard, showcasing its advantages in emotion prediction tasks [29]. While multimodal sentiment analysis has shown promise, several significant challenges persist in the field, including understanding the impact of modalities across datasets and assessing the generalization ability of multimodal sentiment classifiers [24]. Centered on multimodal sentiment analysis, this research delves into the fusion of text, visual, and audio data to provide a nuanced understanding of video sentiments, effectively addressing associated challenges. The synthesis of these three data streams culminates in a unified output, achieved through text analysis, facial expression analysis, and audio analysis.

C. Text Analysis

Sentiment analysis, also known as opinion mining, is a specific task within text analysis that focuses on determining the sentiment or emotional tone expressed in a piece of text [30]–[33]. Sentiment analysis for social media has also grown in popularity [34]. In fact, various techniques have been leveraged to improve both accuracy and interpretability of various sentiment analysis algorithms [35]–[38]. Sentiment analysis of text contents within video content adds complexity to sentiment assessment, involving transcribing spoken language or extracting sentiment from text captions, comments, or subtitles. Sentiment analysis, a subfield of NLP, relies on essential NLP techniques for language analysis [39]–[42]. These techniques are employed to extract sentiments from textual information, which is important when assessing videos that contain a combination of visual and textual elements. In videos, textual content can come from various sources, including speeches, captions, or overlaid text on video frames. Initially, sentiment analysis focused on text analysis, assuming words alone could capture dynamic human sentiment. However, researchers later acknowledged the limitations of words alone in expressing sentiment. Nonetheless, the textual modality remains accessible for deciphering 'positive' and 'negative' elements, employing methods such as learning-based [33] and non-learning-based approaches [32].

D. Facial Expression Recognition

Facial expression analysis is crucial for video sentiment analysis, involving the assessment of facial movement and identification of expressions [43]. Advanced computer vision techniques, like Convolutional Neural Networks (CNNs), enable real-time emotion detection in video frames, particularly important for identifying distressing content, especially for children. Facial expressions are a universal means of communicating emotional states and intentions [44], making them essential for sentiment analysis. While there are attempts to recognize various expressions [45], our focus remains on basic human facial expressions to classify sentiment as ”positive” or ”negative”.

Approximately 55% of emotional information is visual, as facial changes during communication are the first signs of transmitting emotional states [46]. Traditional methods relied on geometric and texture features, but they had limitations due to variations in head pose and demographic constraints like age affecting muscle mobility. Deep learning techniques, such as CNNs, greatly improved accuracy and learning rate.

We explored the usage of CNN with ConvNet, as presented by Gunawan et al. [47]. Individual frames were extracted from the input video and pre-processed through grayscale conversion and histogram equalization to enhance contrast. CNN allowed extraction of numerous facial features for comparison with a database. Additionally, image cropping and resizing were employed to reduce processing time.

E. Audio Analysis

Audio analysis complements visual cues in video sentiment analysis, capturing emotional information conveyed through speech prosody, tone, and acoustic features. Recent advances in deep learning, such as Long Short-Term Memory networks (LSTMs) and CNNs, enhance the accuracy of audio- based sentiment analysis, particularly relevant in child safety scenarios. Music, being an auditory form, significantly influences emotions and plays a crucial role in delivering messages. Analyzing a video's audio is essential, and the Valence-Arousal (V-A) model, popularized by Liu et al., stands out to predict the emotions of songs [48]. This 2-dimensional model classifies emotions into Valence (positive or negative) and Arousal (emotion intensity). Utilizing a Convolutional Neural Network (CNN) on spectrogram representations, their research demonstrated superior performance compared to Support Vector Machines (SVM) [48].

We discovered an article titled 'Predicting the music mood of a song with deep learning' that addressed a multi-classification problem using the Spotify Developer API and Keras Classifier [49]. The author defined four mood labels (energetic, calm, happy, and sad) based on Robert Thayer's traditional mood model [50]. Using Spotify playlists with these mood keywords, the author compiled a dataset of 800 tracks with 10 features from Spotify's 'Get Audio Features' API. The resulting model achieved 76% accuracy, excelling in classifying sad and calm songs.

F. Challenges and Ethical Considerations

Despite rapid advancements, challenges persist in video sentiment analysis. Achieving high accuracy in multimodal sentiment analysis remains complex, demanding the fusion of data from diverse sources. Ensuring fairness, transparency, and accountability in sentiment analysis algorithms, particularly in content filtering for children, raises ethical dilemmas. Striking a balance between free expression and child safety is an ongoing challenge. Finally, akin to other methods of data collection, the utilization of social media data can raise significant ethical issues, encompassing questions about whether this data is deemed private or public and the importance of informed consent.

Another challenge would be the lack of labelled videos. As such, our dataset (text, facial expression, and audio) are from various sources which is publicly available, which will be elaborated in the next section.

III. Datasets

With limited labeled video data, we strategically curated individual datasets for each modality—textual content, facial expressions, and audio. We ensured diversity and relevance by gathering data from various online platforms.

In total, we collected five distinct datasets, each corresponding to one of the three modalities. Subsequent subsections provide detailed exploration, including data sources and pre-processing steps, ensuring data quality and alignment with our research objectives.

A. Text Datasets

We have gathered a total of three distinct labeled text datasets, each sourced from different domains, including platforms like Twitter and text messages. This deliberate approach was taken to broaden the scope of our data collection, ensuring that we encompass a diverse range of textual content for our research.

- Emotions in Text [51]
  This dataset is drawn from the Kaggle open-source community and comprises 21,405 records of texts labelled with emotions. Emotion labels include happiness, sadness, anger, fear, and love.
- Tweet Emotion [52]
  This dataset is drawn from Kaggle open-source community and contains 39,827 records with text from tweets labelled with emotions such as neutral, worry, happiness, sadness, and love.
- Sentiment140 [53]
  This dataset is for academic purposes only and originated as a class project from Stanford University. It comprises 498 records of text labelled with sentiments such as negative, neutral, and positive.

B. Facial Expression Analysis Dataset

The dataset originated from a Kaggle Research Prediction Competition, comprising of grayscale images of faces, each measuring 48x48 pixels [54]. These facial images have undergone an automated registration process to ensure that the face is approximately centered within each image and occupies a consistent amount of space, providing uniformity in the data. The dataset encompasses a total of 35,887 records, categorized into seven distinct emotion classes: “Angry,” “Disgust,” “Fear,” “Happy”, “Sad”, “Surprise”, and “Neutral”.

Notably, the “Disgust” expression class is characterized by a smaller representation, comprising only 600 images, whereas the remaining labels, “Angry”, “Fear”, “Happy”, “Sad”, “Surprise”, and “Neutral”, each contain nearly 5,000 samples. Our main goal for this dataset would be to categorize each face into the labelled categories.

C. Audio Analysis Dataset

We chose to create our own audio dataset for sentiment analysis, as existing datasets mostly contained outdated audio pre-2000s, while TikTok videos predominantly feature modern pop music. Using Spotify playlists labeled with emotions like “Angry”, “Depression”, “Happy”, and ”Calm”, we categorized sentiments as “Positive (Child-friendly)” and “Negative (Potential Harmful)” to align with our research objectives. Utilizing the Spotify API for audio features, we assigned emotions to songs, ensuring dataset integrity through a meticulous review that removed duplicate entries across playlists. After this thorough cleaning and curation process, our final dataset boasts a collection of 788 records. Each entry is not only tagged with its associated emotion but also accompanied by ten distinctive audio features [55]. These features encompass a wide spectrum of musical characteristics, each integral to the intricate world of sentiment in music. These meticulously selected audio features serve as the cornerstone for our sentiment prediction models. They empower us to decode the nuanced emotional landscapes concealed within music, underpinning our mission of precise sentiment analysis.

IV. The Proposed Video Sentiment Analysis Method

A. Overall Methodology

This section provides an overview of the key steps and techniques utilized in our methodology. The proposed multimodal method in this research encompasses various stages and techniques, aimed at assessing video sentiment effectively. We integrate three key components to assess video sentiment as depicted in Fig. 1. The summary of these components is as follows.

Textual Analysis: We transcribe spoken language and extract sentiment from text elements such as captions, comments, and subtitles within the video.

Facial Expression Analysis: We utilize advanced computer vision techniques, including CNNs, to identify and categorize facial expressions within video frames.

Audio Analysis: To complement visual cues, we analyze audio content, emphasizing speech prosody, tone, and acoustic features.

As shown in Fig. 1, our core methodology integrates results from individual component analyses, aggregating three components to classify video sentiment into two categories: positive (Child-friendly) or negative (Potentially Harmful). Individual classifiers are trained for each video component, and the most effective models for each task are chosen. These selected models collaborate in a voting process to determine the overall sentiment of the video.

B. Text Sentiment Analysis

We performed exploratory data analysis on each text dataset, ensuring no empty rows, understanding dataset characteristics, evaluating data distribution, and removing redundant columns like identifiers. We standardized column names to ’content’ and ’sentiment’. After this standardization, we merged the three datasets, resulting in a total of 61,730 rows.

There was a total of 15 unique labels which naturally emerged from the diverse sources of our datasets. There were also similar labels such as “happiness” and “happy”, and also neutral labels. To be consistent with our goal of classifying content as either positive or negative, we removed the neutral labels, and mapped the rest of the labels to be either positive or negative. For clarity, we categorized the following labels as positive: “Enthusiasm”, “Fun”, “Happiness”, “Happy”, “Love”, “Positive”, and “Relief”. Conversely, we categorized the following labels as negative: “Anger”, “Boredom”, “Empty”, “Hate”, “Negative”, “Sadness”, “Worry”, and “Fear”.

After revisiting the count of each label, it was revealed that approximately 59.72% of the labels fell within the negative category, while the remaining 40.28% were categorized as positive. While the distribution did not precisely align with a 50-50 split, it was determined that the dataset did not raise significant concerns of imbalance that could potentially impact the validity of our analysis.

In our text sentiment analysis approach, we followed a systematic methodology. After thorough data preprocessing and cleaning, we conducted label encoding and partitioned the dataset into training, testing, and validation sets. To transform the text data into a numerical format, we utilized Count Vectorizer. Subsequently, we trained a diverse array of models, spanning from simpler ones like Naive Bayes to more intricate ensemble models such as the Voting Classifier. This comprehensive exploration enabled a thorough assessment of the performance of different models in the context of text sentiment analysis.

Furthermore, in pursuit of optimizing our models’ performance, we conducted extensive hyperparameter tuning. This involved leveraging GridSearchCV to systematically identify the most optimal hyperparameters, ensuring that our models were fine-tuned to achieve the best possible results in sentiment analysis. The results of our experiments will be further discussed in Section V.

C. Facial Expression Analysis

There was not much cleaning or data preparation required for the dataset, as there were no missing data and the pixel columns already being represented in a numerical format, with each value in the cell representing a pixel in an image. To be consistent with our overall methodology, we also removed “surprised” and “neutral” labels, which cannot be easily and accurately labelled as “positive” or “negative” sentiment.

In our research, we found that CNNs are commonly employed for facial expression recognition. Hence, we utilized a 7-layer CNN model architecture, which includes the input and output layers. In this architecture, all layers, except for the output layer, utilize the Rectified Linear Unit (ReLU) activation function. The output layer, on the other hand, employs the SoftMax activation function. To align with the emotions represented in the FER-2013 dataset, our output layer consists of 5 nodes, corresponding to the 5 different emotions (after removing neutral and surprised label).

For our model's hyperparameters, we chose the widely-used Adam optimizer, known for its adaptive learning rate capabilities. We set the learning rate to 0.0005, carefully selected after a series of experiments and fine-tuning to strike a balance between achieving rapid convergence and ensuring training stability. Given that facial expression recognition involves multi-class classification, we employed the categorical cross-entropy loss function. This choice facilitates the model's ability to minimize this loss during training, ultimately enhancing its proficiency in accurately classifying facial expressions across multiple categories.

For dataset partitioning, we divided our dataset into two distinct subsets: a training set encompassing 90% of the data and a test set containing the remaining 10%. To maintain a proportionate representation of the various emotion categories, we applied stratification during the dataset split, ensuring that each class was adequately represented in both subsets.

During the model's training process, we opted for a total of 80 epochs. Within each epoch, the model underwent an iteration count equivalent to the length of the training dataset divided by 64. This configuration resulted in approximately 361 steps per epoch. This approach allowed our model to undergo extensive learning and refinement, thereby improving its capability to recognize facial expressions effectively. The corresponding results will be explained and discussed in Section V.

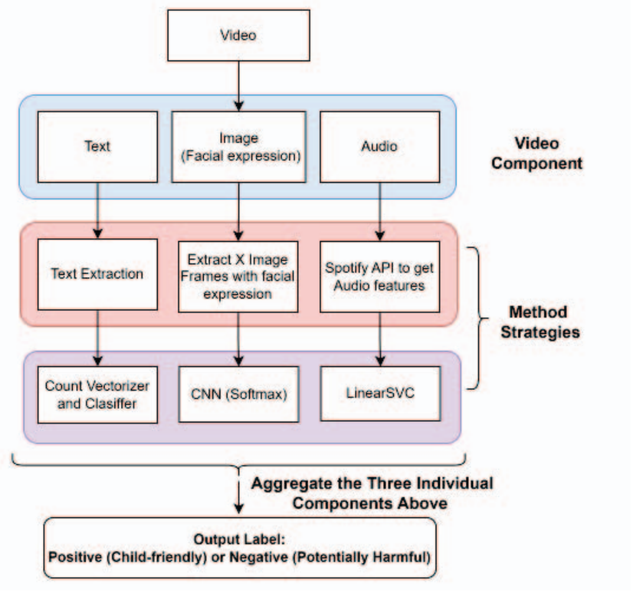
Video
Text
Image (Facial expression)
Audio
Video Component
Text Extraction
Extract X Image Frames with facial expression
Spotify API to get Audio features
Method Strategies
Count Vectorizer and Clasiffer
CNN (Softmax)
LinearSVC
Aggregate the Three Individual Components Above
Output Label:
Positive (Child-friendly) or Negative (Potentially Harmful)

Fig. 1. Architecture overview of proposed method

D. Audio Sentiment Analysis

As part of our exploratory data analysis, we employed violin plots to visualize each feature's distributions, revealing that most features were non-normal. Notably, features like "liveness" displayed positive skewness, while some features like "instrumentalness" exhibited a bimodal distribution. These observations significantly influenced our subsequent data preprocessing decisions, particularly regarding scaling and feature selection. Due to the non-normal distributions observed, we opted for the MinMaxScaler, which preserves the shape of the original distribution while normalizing the data. We proceeded to create an 80-20 train-test split of the dataset, facilitating model evaluation. We also performed label encoding on the labels, ensuring that the labels are mapped to "Positive" and "Negative" correctly and consistent with our work. Specifically, Positive label includes: "Happy" and "Calm" while Negative label includes: "Angry", "Depression". Lastly, after data cleaning and preparation, we began to train our models, with our training strategies encompassing three key phases. The detailed experiment setup is described in Subsection IV-E below.

E. Experiment Setup

The experiment setup for this research primarily focuses on evaluating individual modalities, namely text analysis, facial expression analysis, and audio analysis, with the objective of selecting the most effective model for each component. Our experiments on the individual components will be evaluated mainly using the F1 score, for ensuring the accuracy of sentiment classification. Precision, in particular, is also looked at due to the potential consequences of false positives. For instance, misclassifying potentially harmful content as 'Child-friendly' videos can have serious implications, including safety concerns. To detail our experimental process, we train separate classifiers for text, facial expression, and audio analysis, optimizing each individually through fine-tuning and model selection. Our training strategies encompass traditional, ensemble, and deep learning models if simpler models prove insufficient. In the final phase, a collaborative voting process among selected models determines the overall sentiment of the video, categorizing it as 'Child-friendly' or 'Potentially Harmful'. This approach ensures the final sentiment classification benefits from the strengths of each modality, minimizing the risk of false positives. The experiment's success relies on achieving good performance for each component and ensuring the fusion system accurately provides the final classification output of the video, crucial for safety considerations to prevent mislabeling 'Potentially Harmful' content as 'Child-friendly'. Comprehensive experiment results and discussions are presented in Section V.

V. Results and Discussion

This section evaluates the performance of our proposed method, starting with the validation of individual components—text sentiment analysis, facial expression analysis, and audio sentiment analysis. Subsequently, we integrate these models to assess the overall sentiment of unseen videos. To maintain consistency, we employ four key evaluation metrics: Accuracy, Precision, Recall, and F1-Score, ensuring a comprehensive assessment across all components.

A. Text Sentiment Analysis

In total, we trained several models, spanning from traditional approaches like Naive Bayes and Logistic Regression to ensemble models such as Random Forest, Gradient Boosting, and Voting Classifier. The results are shown in Table I. Based on our results, the best two models are Linear SVC and the Voting Classifier, with the Voting Classifier achieving a slightly better test F1-Score of 85.93%, and Linear SVC at 85.71%.

However, it's crucial to note that Linear SVC exhibits a slightly higher precision rate of 82.50%, which is more than the marginal difference in F1 scores, in contrast to Voting Classifier's precision of 81.30%. Since our primary objective is centered on child safety and reducing the exposure to negative videos, a focus on minimizing false positives can be considered too. Therefore, considering the Precision metric is also pertinent in our decision-making process.

Furthermore, Linear SVC is notably simpler in comparison to the Voting Classifier. Considering these factors, we propose and opt for Linear SVC as the preferred model for this text sentiment analysis component. The accuracy scores are summarised in Table I.

TABLE I
PERFORMANCE EVALUATION OF TEXT SENTIMENT ANALYSIS

Model	Four Evaluation Metrics (%)			
	Accuracy	*Precision*	*Recall*	*F1-Score*
Random Forest	77.62	74.49	**94.62**	83.36
Naive Bayes	80.42	80.40	88.53	84.27
Gradient Boosting	81.05	80.29	90.14	84.93
Linear SVC	**82.39**	**82.50**	89.18	85.71
Voting Classifier	82.33	81.30	91.12	**85.93**

B. Facial Expression Analysis

About facial expression analysis, a 7-layer CNN model is leveraged for this task. The hyperparameters, described in Section IV-C, include the use of the Adam Optimizer, a learning rate of 0.0005, and cross-entropy as the loss function. The model runs over 80 epochs. The performance is summarized in Table II.

TABLE II
PERFORMANCE EVALUATION OF SENTIMENT ANALYSIS ON FACIAL EXPRESSION

Model	Four Evaluation Metrics (%)			
	Accuracy	*Precision*	*Recall*	*F1-Score*
CNN	66.02	67.00	64.96	65.95

The outputs of this facial expression analysis component are not ideal, as the CNN model achieved an F1 score of 65.95% and accuracy of 66.02%. However, such individual performance does not have a negative influence on the final output of the system. This discovery is consistent with previous work, where several weaker individual learning-based models work together to produce a stronger ensemble method.

C. Audio Sentiment Analysis

For audio sentiment analysis, we trained 10 different learning-based models, including traditional models like Decision Trees, ensemble models like Random Forest, and deep learning models such as CNN. The results obtained from the top 5 models are shown in Table III.

TABLE III
PERFORMANCE EVALUATION OF AUDIO SENTIMENT ANALYSIS

Model	Four Evaluation Metrics (%)			
	Accuracy	*Precision*	*Recall*	*F1-Score*
Decision Tree	69.62	64.87	68.57	66.67
AdaBoost	76.58	72.00	77.14	74.48
CNN	79.75	75.00	**81.43**	78.08
CatBoost	81.01	77.03	**81.43**	79.17
Linear SVC	**82.28**	**80.00**	80.00	**80.00**

Note that Decision Tree served as the base model for AdaBoost. As observed, Linear SVC outperformed all other models in terms of F1 score, accuracy, and precision. Hence, Linear SVC is selected as the audio analysis component for performing audio sentiment analysis in our proposed video sentiment analysis task.

D. Integrated All the Individual Components

Using the best model from each individual component, we integrated these models to create a comprehensive pipeline for testing on new, unseen videos. Our pipeline was designed to allow for end-to-end video analysis through a single upload using a web application. The uploaded video undergoes a series of processing steps, including frame extraction to generate images and separation of audio from the video. The audio, which typically includes speeches, is then converted to text. Furthermore, the video's audio information, including music, is extracted for audio sentiment analysis, utilizing the features returned by the Spotify API, as explained earlier.

Each component in the video analysis pipeline produces a binary result of either 'positive' or 'negative,' unless the component is not detected in the video. For instance, if a video lacks background music, the music sentiment component will not return a sentiment. Therefore, it was crucial for us to devise a strategy for testing and validating our overall pipeline.

To evaluate our pipeline, we collected sample videos from TikTok and manually labeled the sentiment for each video. In total, we obtained 60 TikTok videos and labeled them. Due to the multimodal nature of the pipeline, results from all the models are combined and subjected to a voting mechanism to determine a final predicted label.

As shown in Table IV., the overall F1-Score is 73.85%, Precision is 85.50%, and Recall is 65.00%. This reflects the presence of mixed output in some cases. When a video contains both positive and negative predictions from individual component outputs, our predicted label is ’mixed sentiment’ instead of ’positive’ or ’negative.’ This approach accounts for the complexity of mixed sentiments in videos and avoids simplistic binary classifications.

TABLE IV
PERFORMANCE OF INTEGRATED MODEL ON TEST DATA

Model	Four Evaluation Metrics (%)			
	Accuracy	*Precision*	*Recall*	*F1-Score*
Integrated Model	76.67	85.50	65.00	73.85

VI. Conclusion, Limitation, Future Works

A. Conclusion

The surge in online video content has brought sentiment analysis to the forefront, especially in the context of protecting children from harmful material. In response to this crucial need, our research proposes a new multimodal analysis method for assessing video sentiment. The proposed method classifies video sentiment as either positive (child-friendly) or negative (potentially harmful), leveraging insights from three fundamental dimensions: text, facial expressions, and audio.

The evaluation results of our method underscore its practicality and effectiveness in assessing video sentiment, positioning it as a valuable tool for a wide range of applications. By providing a comprehensive overview of the field, our research empowers fellow researchers and practitioners to chart the course of ongoing advancements in video sentiment analysis.

Furthermore, our work has illuminated the current state of the art, shedding light on recent advancements, and charting the path for future directions in video sentiment analysis. As technology continues to evolve and our understanding of sentiment deepens, we foresee a future where video sentiment analysis becomes an even more potent instrument for safeguarding the digital well-being of the younger generation.

B. Limitations

Our methodology involves using Count Vectorizer to transform text into vectors, but it may struggle with multilingual content in social media. Analyzing the presence of sad audio alone may not provide conclusive results for identifying harmful videos, as sad music does not necessarily indicate a negative video (e.g., harmful to children).

In music sentiment analysis, engineered features like danceability are derived using Spotify’s API. However, this approach is limited to songs in Spotify’s database, excluding custom-made or modified music. Our models, developed from scratch, may lack robustness, suggesting potential exploration of transfer learning in the future.

Our research faced challenges due to the scarcity of comprehensive video datasets tailored for sentiment analysis, particularly for children. The limited availability hampers the training and evaluation of models within our specific domain, impacting their generalizability.

Additionally, in the text analysis component, the reliance on external textual sources introduces a potential disconnection between extracted text and actual spoken or presented content, posing a risk to sentiment analysis accuracy. Similarly, the audio dataset used for sentiment analysis is not directly extracted from video audio tracks but sourced from diverse origins, potentially misaligning with nuanced emotional cues specific to our video domain.

C. Future Work

Despite the limitations, our research emphasizes the need for future efforts in curating specialized video datasets that closely align with the detection of child-friendly contents.

Future enhancements in sentiment analysis of online video content can focus on incorporating multilingual support. This involves developing models and techniques to effectively handle and analyze text and speech in multiple languages. Additionally, performing video sentiment analysis on a scale from “Very Positive” to “Very Negative” can provide valuable insights. For example, sentiment labels like “Very Positive” (child-friendly very much), “Positive” (child-friendly), “Neutral,” “Negative” (potentially harmful), and “Very Negative” (very harmful) can be used. Leveraging transfer learning, this future work is expected to achieve exceptional performance while addressing limitations such as the absence of multilingual support. Lastly, explainable sentiment analysis can also be considered in future works.

Acknowledgments

The authors sincerely appreciate the contributions of SMU students Jowett Chng Kai Cheng, Teo Wei Hern, Richard Goh Jiang Wei, and Khong Farn Mun Benjamin to this research. Additionally, the authors thank Dr. Ho Seng-Beng, a senior scientist from IHPC, A*STAR, for valuable discussions.

References

[1] M. Yesilada and S. Lewandowsky, “Systematic review: Youtube recommendations and problematic content,” *Internet policy review*, vol. 11, no. 1, 2022.

[2] E. Abi-Jaoude, K. T. Naylor, and A. Pignatiello, “Smartphones, social media use and youth mental health,” *Cmaj*, vol. 192, no. 6, pp. E136–E141, 2020.

[3] L. Stappen, A. Baird, E. Cambria, and B. W. Schuller, “Sentiment analysis and topic recognition in video transcriptions,” *IEEE Intelligent Systems*, vol. 36, no. 2, pp. 88–95, 2021.

[4] L.-P. Morency, R. Mihalcea, and P. Doshi, “Towards multimodal sentiment analysis: Harvesting opinions from the web,” in *Proceedings of the 13th international conference on multimodal interfaces*, 2011, pp. 169–176.

[5] M. Soleymani, D. Garcia, B. Jou, B. Schuller, S.-F. Chang, and M. Pantic, “A survey of multimodal sentiment analysis,” *Image and Vision Computing*, vol. 65, pp. 3–14, 2017.

[6] D. Rajagopal, E. Cambria, D. Olsher, and K. Kwok, “A graph-based approach to commonsense concept extraction and semantic similarity detection,” in *WWW*, 2013, pp. 565–570.

[7] E. Cambria and A. Hussain, “Sentic album: Content-, concept-, and context-based online personal photo management system,” *Cognitive Computation*, vol. 4, no. 4, pp. 477–496, 2012.

[8] E. Cambria, A. Hussain, C. Havasi, and C. Eckl, “Common sense computing: From the society of mind to digital intuition and beyond,” in *Biometric ID Management and Multimodal Communication*, ser. Lecture Notes in Computer Science. Berlin Heidelberg: Springer, 2009, vol. 5707, pp. 252–259.

[9] E. Cambria, S. Poria, F. Bisio, R. Bajpai, and I. Chaturvedi, “The CLSA model: A novel framework for concept-level sentiment analysis,” in *LNCS*. Springer, 2015, vol. 9042, pp. 3–22.

[10] E. Cambria, T. Mazzocco, A. Hussain, and C. Eckl, “Sentic medoids: Organizing affective common sense knowledge in a multi-dimensional vector space,” ser. Lecture Notes in Computer Science. Berlin Heidelberg: Springer-Verlag, 2011, vol. 6677, pp. 601–610.

[11] K. Zhang, Y. Li, J. Wang, E. Cambria, and X. Li, “Real-time video emotion recognition based on reinforcement learning and domain knowledge,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 3, pp. 1034–1047, 2021.

[12] A. Valdivia, V. Luzón, E. Cambria, and F. Herrera, “Consensus vote models for detecting and filtering neutrality in sentiment analysis,” *Information Fusion*, vol. 44, pp. 126–135, 2018.

[13] E. Cambria, R. Mao, M. Chen, Z. Wang, and S.-B. Ho, “Seven pillars for the future of AI,” *IEEE Intelligent Systems*, vol. 38, no. 6, 2023.

[14] S. Cavallari, E. Cambria, H. Cai, K. Chang, and V. Zheng, “Embedding both finite and infinite communities on graph,” *IEEE Computational Intelligence Magazine*, vol. 14, no. 3, pp. 39–50, 2019.

[15] E. Cambria, P. Chandra, A. Sharma, and A. Hussain, “Do not feel the trolls,” in *ISWC*, Shanghai, 2010.

[16] F. Xing, E. Cambria, and R. Welsch, “Intelligent asset allocation via market sentiment views,” *IEEE Computational Intelligence Magazine*, vol. 13, no. 4, pp. 25–34, 2018.

[17] E. Cambria, T. Benson, C. Eckl, and A. Hussain, “Sentic PROMs: Application of sentic computing to the development of a novel unified framework for measuring health-care quality,” *Expert Systems with Applications*, vol. 39, no. 12, pp. 10 533–10 543, 2012.

[18] H. T. Nguyen, P. H. Duong, and E. Cambria, “Learning short-text semantic similarity with word embeddings and external knowledge sources,” *Knowledge-Based Systems*, vol. 182, no. 104842, 2019.

[19] E. Cambria, N. Howard, J. Hsu, and A. Hussain, “Sentic blending: Scalable multimodal fusion for continuous interpretation of semantics and sentics,” in *IEEE SSCI*, Singapore, 2013, pp. 108–117.

[20] M. Abbas, K. A. Memon, A. A. Jamali, S. Memon, and A. Ahmed, “Multinomial naive bayes classification model for sentiment analysis,” *IJCSNS Int. J. Comput. Sci. Netw. Secur*, vol. 19, no. 3, p. 62, 2019.

[21] T. Wu, J. Peng, W. Zhang, H. Zhang, S. Tan, F. Yi, C. Ma, and Y. Huang, “Video sentiment analysis with bimodal information-augmented multi-head attention,” *Knowledge-Based Systems*, vol. 235, p. 107676, 2022.

[22] L. Stappen, A. Baird, E. Cambria, and B. W. Schuller, “Sentiment analysis and topic recognition in video transcriptions,” *IEEE Intelligent Systems*, vol. 36, no. 2, pp. 88–95, 2021.

[23] S. A. Abdu, A. H. Yousef, and A. Salem, “Multimodal video sentiment analysis using deep learning approaches, a survey,” *Information Fusion*, vol. 76, pp. 204–226, 2021.

[24] S. Poria, N. Majumder, D. Hazarika, E. Cambria, A. Gelbukh, and A. Hussain, “Multimodal sentiment analysis: Addressing key issues and setting up the baselines,” *IEEE Intelligent Systems*, vol. 33, no. 6, pp. 17–25, 2018.

[25] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, “A review of affective computing: From unimodal analysis to multimodal fusion,” *Information fusion*, vol. 37, pp. 98–125, 2017.

[26] T. Yue, R. Mao, H. Wang, Z. Hu, and E. Cambria, “KnowleNet: Knowledge fusion network for multimodal sarcasm detection,” *Information Fusion*, vol. 100, p. 101921, 2023.

[27] R. Mao, X. Li, M. Ge, and E. Cambria, “MetaPro: A computational metaphor processing model for text pre-processing,” *Information Fusion*, vol. 86, pp. 30–43, 2022.

[28] X. Wang, L. Zhu, Z. Zheng, M. Xu, and Y. Yang, “Align and tell: Boosting text-video retrieval with local alignment and fine-grained supervision,” *IEEE Transactions on Multimedia*, 2022.

[29] K. Lin, X. Wang, Z. Zheng, L. Zhu, and Y. Yang, “Less is more: Sparse sampling for dense reaction predictions,” *arXiv preprint arXiv:2106.01764*, 2021.

[30] J. Cui, Z. Wang, S.-B. Ho, and E. Cambria, “Survey on sentiment analysis: evolution of research methods and topics,” *Artificial Intelligence Review*, vol. 56, pp. 8469–8510, 2023.

[31] Z. Wang, S.-B. Ho, and E. Cambria, “A review of emotion sensing: categorization models and algorithms,” *Multimedia Tools and Applications*, vol. 79, pp. 35 553–35 582, 2020.

[32] Z. Wang, Z. Hu, F. Li, S.-B. Ho, and E. Cambria, “Learning-based stock trending prediction by incorporating technical indicators and social media sentiment,” *Cognitive Computation*, vol. 15, pp. 1092–1102, 2023.

[33] Z. Wang, Z. Hu, S.-B. Ho, E. Cambria, and A.-H. Tan, “MiMuSA—mimicking human language understanding for fine-grained multi-class sentiment analysis,” *Neural Computing and Applications*, vol. 35, pp. 15 907–15 921, 2023.

[34] L. Oneto, F. Bisio, E. Cambria, and D. Anguita, “Statistical learning theory and ELM for big social data analysis,” *IEEE Computational Intelligence Magazine*, vol. 11, no. 3, pp. 45–55, 2016.

[35] F. Xing, F. Pallucchini, and E. Cambria, “Cognitive-inspired domain adaptation of sentiment lexicons,” *Information Processing and Management*, vol. 56, no. 3, pp. 554–564, 2019.

[36] Z. Wang, S.-B. Ho, and E. Cambria, “Multi-level fine-scaled sentiment sensing with ambivalence handling,” *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 28, no. 4, pp. 683–697, 2020.

[37] D. Vilares, H. Peng, R. Satapathy, and E. Cambria, “BabelSenticNet: A commonsense reasoning framework for multilingual sentiment analysis,” in *IEEE SSCI*, 2018, pp. 1292–1298.

[38] E. Cambria, B. Schuller, B. Liu, H. Wang, and C. Havasi, “Statistical approaches to concept-level sentiment analysis,” *IEEE Intelligent Systems*, vol. 28, no. 3, pp. 6–9, 2013.

[39] Z. Wang, V. Joo, C. Tong, X. Xin, and H. C. Chin, “Anomaly detection through enhanced sentiment analysis on social media data,” in *International Conference on Cloud Computing Technology and Science*. IEEE, 2014, pp. 917–922.

[40] Z. Wang, V. Joo, C. Tong, and D. Chan, “Issues of social data analytics with a new method for sentiment analysis of social media data,” in *International Conference on Cloud Computing Technology and Science*. IEEE, 2014, pp. 899–904.

[41] Z. Wang, V. J. C. Tong, and H. C. Chin, “Enhancing machine-learning methods for sentiment classification of web data,” in *Information Retrieval Technology*. Springer, 2014, pp. 394–405.

[42] E. Cambria, Q. Liu, S. Decherchi, F. Xing, and K. Kwok, “SenticNet 7: A commonsense-based neurosymbolic ai framework for explainable sentiment analysis,” in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 2022, pp. 3829–3839.

[43] Y. Tian, T. Kanade, and J. F. Cohn, “Facial expression recognition,” *Handbook of face recognition*, pp. 487–519, 2011.

[44] S. Li and W. Deng, “Deep facial expression recognition: A survey,” *IEEE transactions on affective computing*, vol. 13, no. 3, pp. 1195–1215, 2020.

[45] V. Bettadapura, “Face expression recognition and analysis: the state of the art,” *arXiv preprint arXiv:1203.6722*, 2012.

[46] W. Mellouk and H. Wahida, “Facial emotion recognition using deep learning: review and insights,” *Procedia Computer Science*, vol. 175, pp. 689–694, 01 2020.

[47] T. S. Gunawan, A. Ashraf, B. S. Riza, E. V. Haryanto, R. Rosnelly, M. Kartiwi, and Z. Janin, “Development of video-based emotion recognition using deep learning with google colab,” *TELKOMNIKA*, vol. 18, no. 5, pp. 2463–2471, 2020.

[48] T. Liu, L. Han, L. Ma, and D. Guo, “Audio-based deep music emotion recognition,” in *AIP Conference Proceedings*, vol. 1967, 2018.

[49] C. V, “Predicting the music mood of a song with deep learning.” Sep 2020. [Online]. Available: https://towardsdatascience.com/predicting-the-music-mood-of-a-song-with-deep-learning-c3ac2b45229e

[50] R. E. Thayer, *The biopsychology of mood and arousal*. Oxford University Press, 1990.

[51] “Emotions in text.” [Online]. Available: https://www.kaggle.com/datasets/ishantjuyal/emotions-in-text

[52] “Emotion detection from text.” [Online]. Available: https://www.kaggle.com/datasets/pashupatigupta/emotion-detection-from-text

[53] “Sentiment140.” [Online]. Available: http://help.sentiment140.com/for-students

[54] C. Pramerdorfer and M. Kampel, “Facial expression recognition using convolutional neural networks: state of the art,” *arXiv preprint arXiv:1612.02903*, 2016.

[55] R. Panda, R. M. Malheiro, and R. P. Paiva, “Audio features for music emotion recognition: a survey,” *IEEE Transactions on Affective Computing*, 2020.