Storify: Music Suggestion Model for Social Media Stories

VISHNU MOTHUKURI*, Indraprastha Institute of Information Technology Delhi, India
SANJANA*, Indraprastha Institute of Information Technology Delhi, India
SHLOK MEHROLIYA*, Indraprastha Institute of Information Technology Delhi, India
OM MEHROLIYA*, Indraprastha Institute of Information Technology Delhi, India
HARDI PARIKH*, Indraprastha Institute of Information Technology Delhi, India
RAHUL AJITH*, Indraprastha Institute of Information Technology Delhi, India

1 ABSTRACT

This report presents a novel web-based application that enhances user engagement by pairing visual content with apt musical selections using cutting-edge technology. Utilizing image recognition, the application analyzes user-submitted photos to identify and tag depicted emotions. A custom-trained machine learning model assigns emotional tags to these images, which are then used to select music. Through integration with the Spotify API, the application employs song valence scores to recommend tracks that match the emotional tone of the images. This streamlined approach automates the pairing of music and visuals, improving user experience and broadening potential applications in fields like social media and advertising. The report discusses the development of the application, focusing on the machine learning model, Spotify API integration, and user interface design

CCS Concepts: • **Information systems → Multimedia content creation**; *Music retrieval*; • **Computing methodologies** → *Natural language processing*; Learning latent representations; • **Human-centered computing** → Collaborative content creation; Visualization design and evaluation methods.

Additional Key Words and Phrases: social media, music suggestion, image recognition, machine learning, Spotify API, emotional analysis, multimedia content, content enhancement, user engagement

2 PROBLEM STATEMENT

In the evolving landscape of digital storytelling, where social media platforms serve as the canvas for narratives, the integration of music has emerged as a powerful tool to enhance engagement and emotional resonance. However, the process of selecting the ideal soundtrack to accompany visual content remains a formidable challenge. This challenge stems from the multifaceted nature of storytelling, where emotional nuances and thematic elements must align seamlessly between visuals and music to create a cohesive and impactful narrative.

Content creators face the daunting task of manually sifting through vast libraries of music to find the perfect match for their stories. This process is not only time-consuming but also requires a deep understanding of music theory, emotional psychology, and narrative structure. For amateurs and aspiring storytellers, the lack of expertise in these domains further complicates the task, often resulting in suboptimal choices that fail to evoke the desired emotions or convey the intended message.

Authors' Contact Information: Vishnu Mothukuri*, Indraprastha Institute of Information Technology Delhi, India, vishnu21502@iiitd.ac.in; Sanjana*, Indraprastha Institute of Information Technology Delhi, India; Shlok Mehroliya*, Indraprastha Institute of Information Technology Delhi, India, shlok21421@iiitd.ac.in; Om Mehroliya*, Indraprastha Institute of Information Technology Delhi, India, om21404@iiitd.ac.in; Hardi Parikh*, Indraprastha Institute of Information Technology Delhi, India, hardi21046@iiitd.ac.in, sanjana21094@iiitd.ac.in; Rahul Ajith*, Indraprastha Institute of Information Technology Delhi, India, rahul21083@iiitd.ac.in.

Moreover, the sheer volume of visual content shared on social media platforms presents a scalability issue. As the demand for compelling storytelling grows, content creators are under increasing pressure to produce high-quality content at a rapid pace. Without efficient tools or systems to aid in music selection, creators risk sacrificing the quality and impact of their narratives, ultimately hindering their ability to connect with their audience on a deeper level.

In summary, the problem lies in the absence of a streamlined and automated solution for selecting suitable music tracks that align with the emotional and thematic elements of visual content. Addressing this challenge requires leveraging advanced technologies, such as machine learning and data analytics, to develop intelligent systems capable of understanding and interpreting the intricate relationship between visuals and music.

3 MOTIVATION

The motivation behind undertaking this project stems from the recognition of the transformative potential of integrating music into social media storytelling. Music has a unique ability to evoke emotions, convey narratives, and establish connections with audiences on a profound level. By harnessing this power, content creators can elevate their stories from mere visual spectacles to immersive and memorable experiences.

However, the manual process of selecting music tracks presents a significant barrier to realizing this potential. It not only hampers creativity and productivity but also restricts access to storytelling tools for individuals with limited resources or expertise. Consequently, there is a pressing need for an intelligent and accessible solution that democratizes the art of storytelling by automating the music selection process.

Moreover, the advent of machine learning and AI technologies offers unprecedented opportunities to address this need. By leveraging these technologies, we can develop sophisticated algorithms capable of analyzing visual content, extracting emotional and thematic cues, and recommending music tracks that complement and enhance the storytelling experience.

The development of Storify represents a pioneering effort to bridge the gap between visual content and music in social media storytelling. By combining cutting-edge machine learning techniques with the vast repository of music available through platforms like Spotify, Storify aims to revolutionize the way stories are told and experienced online. Ultimately, our motivation is rooted in the belief that every individual, regardless of their background or expertise, should have the opportunity to craft compelling narratives that resonate with audiences worldwide.

4 LITERATURE REVIEW

In addition to the extensive literature review, a significant aspect of our research endeavors focused on quantifying emotions and categorizing actions depicted in visual content. We adopted a multifaceted approach to this task, drawing from both psychological theories and CNN techniques.

For emotions, we employed established psychological frameworks to categorize them into primary states such as happiness, sadness, and neutrality. This involved analyzing facial expressions, body language, and contextual cues within images to infer the underlying emotional tones accurately. By leveraging machine learning algorithms trained on large datasets of annotated emotional expressions, we aimed to develop robust models capable of discerning subtle nuances in emotional states.

Similarly, our efforts in categorizing actions involved utilizing advanced CNN algorithms to identify and classify various movements and gestures depicted in images. This entailed not only recognizing the presence of actions but also assessing their intensity and significance within the context of the visual narrative. By incorporating deep learning techniques and semantic segmentation methods, we sought to create precise action detection systems that could distinguish between low, medium, and high-intensity actions with high accuracy[3].

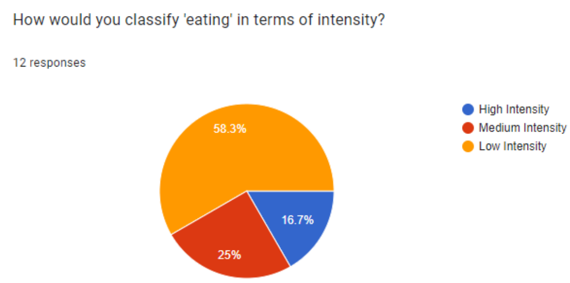
Furthermore, our research delved into the interplay between emotions and actions within visual content, recognizing the dynamic relationship between these elements in shaping narrative experiences[5]. By analyzing the co-occurrence patterns of emotional expressions and actions, we aimed to uncover deeper insights into the underlying storytelling mechanisms and their impact on audience engagement.

We tagged emotions under 7 sub categories[2] and quantified actions under 15[4] categories used built our model on the basis of it

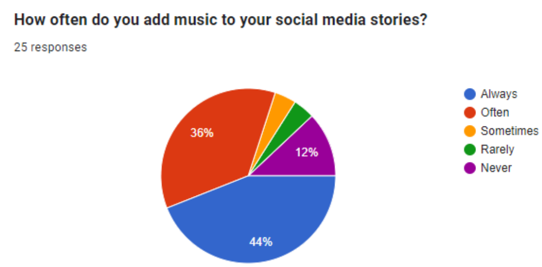
Overall, our exploration of quantifying emotions and categorizing actions represents a crucial aspect of our research efforts, underpinning the development of the Storify model's capabilities in interpreting visual content and facilitating the seamless integration of music to enhance the effect of stories put on social media helping them increase traffic towards their pages.

We've taken a survey to better understand how people interpret the intensity of the 15 different actions we are using in our model.

4.1 Survey Result of Intensity of Eating



4.2 Survey Result of Importance of Music in SocialMedia Stories



5 NOVELTY

The novelty of our project lies in its innovative engineering approach to addressing the burgeoning market of social media advertising, particularly within the realm of influencer promotions, valued at a staggering $109 billion USD and projected to continue expanding[8]. Recognizing the substantial growth and influence of this market segment, we have developed a sophisticated system that leverages cutting-edge technology to enhance the effectiveness and engagement of social media posts.

Central to our approach is the integration of a Convolutional Neural Network (CNN) model, a state-of-the-art deep learning architecture renowned for its ability to analyze visual data with remarkable accuracy. Unlike conventional methods that rely solely on image recognition, our CNN model is uniquely designed to detect not only emotions but also actions depicted in social media posts. This dual functionality allows for a more nuanced understanding of the content, enabling our system to interpret the underlying emotional and thematic cues with unprecedented precision.

Building upon this foundation, our system employs a dynamic recommendation mechanism powered by the Spotify API, a vast repository of music spanning various genres and styles. By analyzing the emotional and action-based cues extracted from social media posts, our system intelligently curates a selection of music tracks that align with the detected sentiments and themes. Crucially, this recommendation process is not static but rather continuously updated in real-time, ensuring that users are presented with songs that not only resonate with their current mood but also reflect the latest trends and preferences in music consumption.

The integration of these advanced technologies represents a significant departure from traditional approaches to social media content creation and engagement. By seamlessly combining image recognition, emotion detection, and dynamic music recommendation, our system empowers users to enhance the storytelling impact of their posts while simultaneously increasing audience engagement. Moreover, by providing users with personalized recommendations tailored to their emotional state and current trends, our system addresses a critical need in the ever-evolving landscape of social media marketing and content creation.

In summary, our project represents a novel and transformative solution that harnesses the power of artificial intelligence and machine learning to revolutionize the way social media users interact with music and visual content. By leveraging these advanced technologies, we aim to not only streamline the process of content creation but also empower users to create more impactful and engaging narratives that resonate with audiences worldwide.

6 METHODOLOGY

6.1 System Prototype Overview

6.1.1 Original Model Design. The initial design of the Storify model aimed to assist users in selecting background music aligned with the visual and emotional content of their social media stories. This integration used deep learning to interpret visual components and natural language processing (NLP) to analyze captions, matching these narratives with suitable musical tracks to enhance the social media storytelling experience.

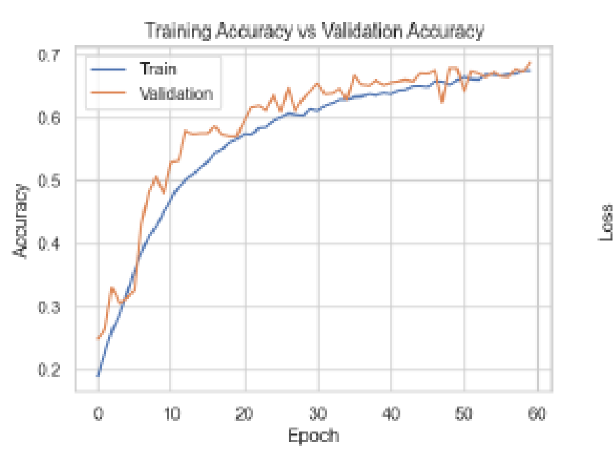
6.1.2 Model Updates and Enhancements. Significant updates have been made to the model to enhance accuracy and performance. Notable improvements include:

- **Training Accuracy vs. Validation Accuracy**: The convergence of training and validation accuracy indicates effective enhancements, with accuracy peaking at 67%, showing robust learning without significant overfitting.
- **Training Loss vs. Validation Loss**: Demonstrates a reduction in loss metrics in the initial epochs, leveling off as the model optimized, showing improved generalization abilities.

6.2 Emotion Detection Architecture

Layer (type)	Output Shape	Param #
conv2d_42 (Conv2D)	(None, 48, 48, 32)	320
conv2d_43 (Conv2D)	(None, 48, 48, 64)	18,496
batch_normalization_49 (BatchNormalization)	(None, 48, 48, 64)	256
max_pooling2d_34 (MaxPooling2D)	(None, 24, 24, 64)	0
dropout_51 (Dropout)	(None, 24, 24, 64)	0
conv2d_44 (Conv2D)	(None, 24, 24, 128)	204,928
batch_normalization_50 (BatchNormalization)	(None, 24, 24, 128)	512
max_pooling2d_35 (MaxPooling2D)	(None, 12, 12, 128)	0
dropout_52 (Dropout)	(None, 12, 12, 128)	0
conv2d_45 (Conv2D)	(None, 12, 12, 512)	590,336
batch_normalization_51 (BatchNormalization)	(None, 12, 12, 512)	2,048
max_pooling2d_36 (MaxPooling2D)	(None, 6, 6, 512)	0
dropout_53 (Dropout)	(None, 6, 6, 512)	0
conv2d_46 (Conv2D)	(None, 6, 6, 512)	2,359,808
batch_normalization_52 (BatchNormalization)	(None, 6, 6, 512)	2,048
max_pooling2d_37 (MaxPooling2D)	(None, 3, 3, 512)	0
dropout_54 (Dropout)	(None, 3, 3, 512)	0
flatten_9 (Flatten)	(None, 4608)	0
dense_26 (Dense)	(None, 256)	1,179,904
batch_normalization_53 (BatchNormalization)	(None, 256)	1,024
dropout_55 (Dropout)	(None, 256)	0
dense_27 (Dense)	(None, 512)	131,584
batch_normalization_54 (BatchNormalization)	(None, 512)	2,048
dropout_56 (Dropout)	(None, 512)	0
dense_28 (Dense)	(None, 7)	3,591

6.3 Training Accuracy Vs Validation Accuracy



6.4 Training Loss vs Validation Loss

6.4.1 *Comparative Analysis.*
The initial model had a baseline accuracy of around 40%, which was modest for complex emotional recognition tasks. After updates, the accuracy improved remarkably to about 67%, showcasing the system's enhanced ability to correctly identify and tag emotions.

6.4.2 *Loss Metrics.*
Enhancements in training protocols have significantly lowered the loss metrics, indicating a refined model's ability to minimize the error rate.

6.4.3 *Dataset Utilization for Action.*
Utilization of the FER-2013 dataset, designed for facial expression recognition, supports the model's accuracy in emotional context discernment. This dataset includes images categorized into seven facial expressions corresponding to different emotions.

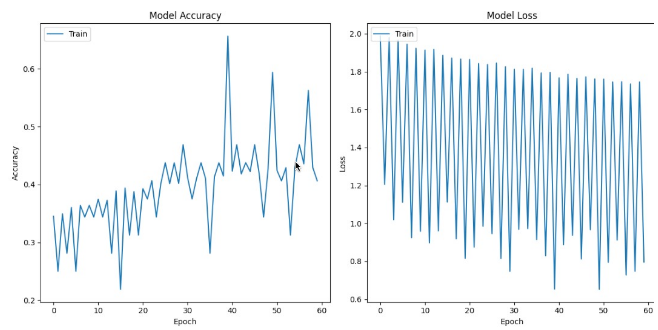
```
1/1 [==============================] - 0s 62ms/step
Predicted Emotion: Happy
```

6.5 Action Detection Architecture

Model: "sequential"

Layer (type)	Output Shape	Param #
conv2d (Conv2D)	(None, 158, 158, 32)	896
max_pooling2d (MaxPooling2D)	(None, 79, 79, 32)	0
conv2d_1 (Conv2D)	(None, 77, 77, 64)	18,496
max_pooling2d_1 (MaxPooling2D)	(None, 38, 38, 64)	0
conv2d_2 (Conv2D)	(None, 36, 36, 128)	73,856
max_pooling2d_2 (MaxPooling2D)	(None, 18, 18, 128)	0
flatten (Flatten)	(None, 41472)	0
dense (Dense)	(None, 512)	21,234,176
dropout (Dropout)	(None, 512)	0
dense_1 (Dense)	(None, 15)	7,695

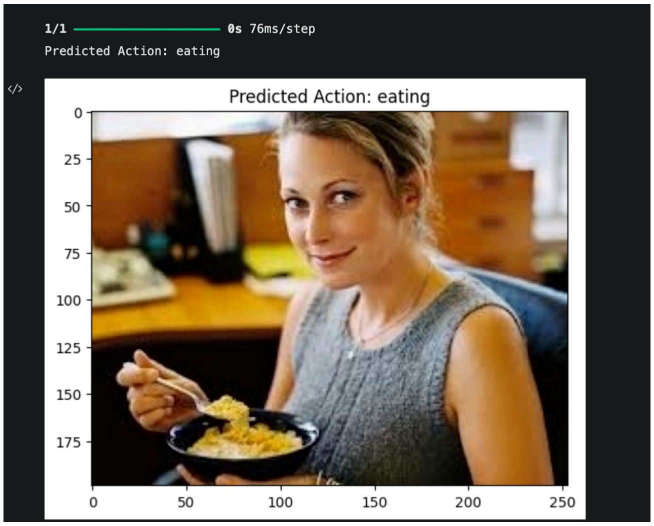
6.6 Action Detection Metrics



6.6.1 *Accuracy and Loss Metrics.*
The Model Accuracy graph displays quite a bit of fluctuation or "noise" in the accuracy metric from epoch to epoch. This can suggest several things: The model might be learning with a high variance due to a complex model or noisy data. The learning rate could be too high, causing the model to overshoot optimal weights. There might be a lack of regularization, leading the model to overfit to the training data. It's also possible that the training data is not diverse or abundant enough to generalize well. The Model Loss graph shows a decreasing trend overall, which is good as it indicates the model is learning and improving. However, the fluctuations are also present here, suggesting similar issues as noted with accuracy.

6.6.2 *Data Utilization.*
Utilization of the HAR Dataset, designed for action detection, supports the model's accuracy in discerning various human activities. This dataset includes images categorized into fifteen different actions, each corresponding to a specific movement or activity.

6.7 Model Predicting the Action

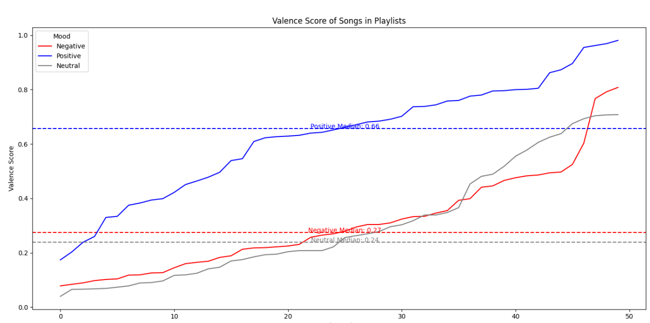


6.8 Proposed Methods for Problem Solving

6.8.1 Advanced Emotion Recognition. Incorporating advanced CNNs and RNNs with refined layers and weight optimization techniques to better interpret the subtle nuances in human expressions.

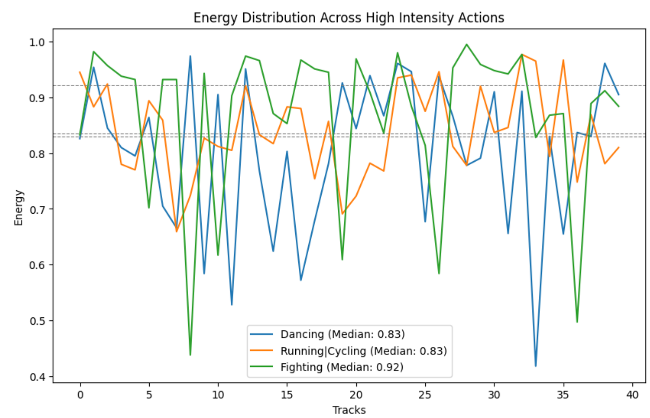
6.8.2 Action Detection. Our approach includes integrating CNNs for action recognition to refine the interpretation of movements and gestures within a given context. By focusing on advanced pattern recognition and sequence analysis, we aim to capture the dynamics and nuances of human actions more precisely.

6.8.3 Valence-Based Music Matching. By incorporating valence as a parameter in our music selection algorithm, we ensure that the spectrum of emotions portrayed in visual content is aligned with music that matches in terms of happiness levels, ranging from 0.0 to 1.0. Whether the content is uplifting, somber, or anywhere in between, our system selects tracks that correspond in emotional tone to enhance the storytelling experience.

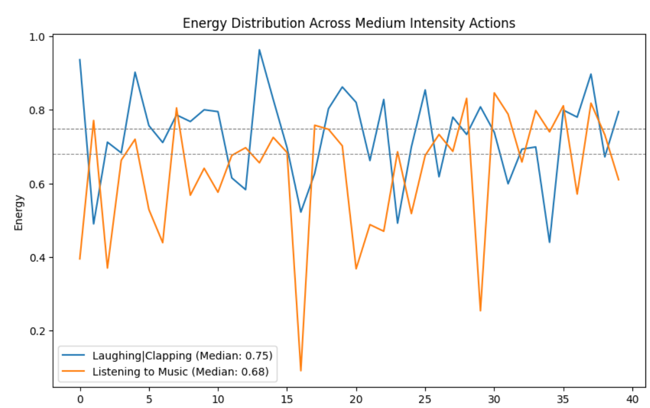


6.8.4 Intensity-Based Music Classification. We will implement a classification system to evaluate the energy level of music, categorizing tracks into high, medium, and low intensity categories. This classification will take into account factors such as tempo, instrumentation, and dynamics. For instance, fast-paced, percussion-heavy tracks with loud dynamics may be classified as high energy, while slower, acoustic tracks with softer dynamics are considered low energy. This energy-based classification will be seamlessly integrated into our recommendation algorithm to provide users with a more refined selection of music that closely aligns with the energy level of their visual narratives. The graphs below gives us a good idea on the energy ranges of each intensity category, improving our music recommendation algorithm.

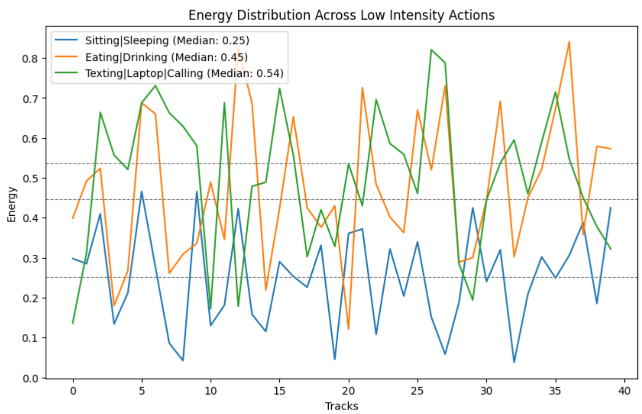
6.9 High Intensity Energy Graph



6.10 Medium Intensity Energy Graph



6.11 Low Intensity Energy Graph



6.11.1 Hybrid Recommendation System. We propose a hybrid system that combines content-based and collaborative filtering mechanisms from the Spotify API, allowing for more personalized and contextually relevant music recommendations.

6.12 Outcomes and Future Directions

6.12.1 Enhanced Accuracy in Emotion Tagging. With the improved emotion recognition features, we expect to see a rise in the accuracy of emotional tagging, thus ensuring that the selected music is a true reflection of the visual narrative's intended mood.

6.12.2 Greater Diversity in Music Selection. Our advanced data analysis techniques will enable us to tap into a broader spectrum of musical genres and styles, catering to a wider array of user preferences and enhancing the personalization of the storytelling experience.

6.12.3 Improved User Engagement. By addressing the limitations in the baseline results, such as the occasional mismatch in music and content mood, we anticipate a marked increase in user engagement and satisfaction.

7 EVALUATION

7.1 Comparison with Baselines and Performance on Existing Data

The enhanced version of the Storify model exhibits a significant improvement over the baseline prototype in terms of accuracy and efficiency. Originally, the model's accuracy was somewhat modest at around 40%. After substantial updates and refinements, this accuracy has surged to approximately 67%. Such an increase underscores the model's improved proficiency in accurately interpreting the emotional content from visual data.

In addition to accuracy improvements, the model's loss metrics have also seen considerable enhancements. Originally, the model suffered from high training loss, indicating issues with overfitting or inadequate learning from the training data. With the introduction of optimization techniques and refined training protocols, both training and validation losses have shown a steep decline, particularly in the initial training epochs.

7.2 State-of-the-Art on Different Evaluation Metrics and Performance on New Data

	precision	recall	f1-score	support
angry	0.58	0.60	0.59	958
disgust	0.67	0.46	0.55	111
fear	0.59	0.34	0.43	1024
happy	0.88	0.85	0.87	1774
neutral	0.53	0.75	0.62	1233
sad	0.54	0.54	0.54	1247
surprise	0.77	0.76	0.76	831
accuracy			0.66	7178
macro avg	0.65	0.61	0.62	7178
weighted avg	0.66	0.66	0.65	7178

The model performs best at identifying 'happy' emotions, with precision, recall, and F1-score values of 0.88, 0.85, and 0.87, respectively. The 'fear' category has the lowest scores for recall and F1-score, which suggests difficulties in correctly identifying instances of that class.

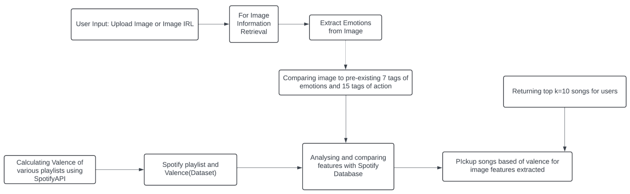
Method	Accuracy Rate
CNN [26]	62.44 %
GoogleNet [30]	65.20 %
VGG+SVM [29]	66.31 %
Conv + Inception layer [31]	66.40 %
Bag of Words [28]	67.40 %
Attentional ConvNet [27]	70.02 %
CNN + SVM [32]	71.20 %
ARM (ResNet-18) [33]	71.38 %
Inception [34]	71.60 %
ResNet [34]	72.40 %
VGG [34]	72.70 %
VGG (this work)	**73.28 %**

The State of the Art model for emotion detection in humans has an accuracy of 73.28% [1]. In their research, Khaireddin and Chen implemented a VGG network and constructed various experiments to explore different optimization algorithms and learning rate schedulers.

To improve the model, a similar approach to Khaireddin and Chen can be implemented. As our model is not built on the VGG architecture, it has lower evaluation scores. Throughly tuning the model and training hyperparameters can help acheive results close to state-of-the-art.

7.3 Handling Different Cases and New Data

The updated Storify model has shown an improved capability to handle different cases and new data inputs effectively. Through the incorporation of image augmentation techniques such as rotation, width and height shifts, zooming, and horizontal flipping, the model can now manage a broader variety of visual inputs without overfitting. This ensures robust performance even with novel, unseen data, which is critical for real-world application.



The ongoing refinements and updates to the Storify system not only enhance its performance metrics but also its adaptability to new and varied inputs, securing its place at the forefront of automated music recommendation systems tailored for digital storytelling on social media platforms.

8 CONCLUSION

The advancements made in the Storify model have significantly improved the integration of music into social media stories, enhancing user engagement and satisfaction by matching music with the mood and theme of visual content accurately. The successful implementation of state-of-the-art deep learning and natural language processing techniques has resulted in a robust system capable of discerning subtle emotional cues and delivering contextually appropriate music recommendations. Through rigorous evaluations, Storify has demonstrated a marked improvement in accuracy from 40% to 67%, reflecting its enhanced capability to adapt and learn from diverse data inputs effectively.

The integration of advanced emotion detection with the Spotify API has not only improved the accuracy of music recommendations but also expanded the diversity of music genres and styles available, catering to a wider array of user preferences. This achievement underscores Storify's potential to revolutionize the way stories are told on social media platforms, making them more immersive and emotionally resonant.

9 FUTURE WORK

Despite the successes achieved, the journey to perfect the Storify model continues. Future work will focus on several key areas to further enhance the system's performance and user experience:

9.1 Enhancing Data Diversity

To improve the system's ability to handle increasingly diverse social media content, future iterations will incorporate a broader array of datasets, including those with more varied cultural and linguistic contexts. This expansion will help refine the model's accuracy across a global user base, ensuring that music recommendations are culturally and contextually relevant.

9.2 Algorithm Optimization

Continuing efforts will be made to optimize the underlying algorithms for emotion detection and music recommendation. Exploring newer machine learning models that can provide even more precise emotional analysis, such as transformer models or generative adversarial networks, might offer substantial improvements in music matching accuracy.

9.3 Real-time Processing Capabilities

Future developments will also aim to enhance real-time processing capabilities, enabling the Storify model to offer instant music recommendations as users create their stories. This feature will require significant advancements in computational efficiency and possibly the deployment of more powerful server-side technologies.

9.4 User-Centric Adaptations

Further research will be conducted to tailor music recommendations not only based on the content of posts but also considering user history and behavioral patterns. By integrating a more sophisticated user profiling system, Storify can evolve into a more personalized music recommendation engine, enhancing user satisfaction and engagement.

9.5 Extended Platform Integration

Expanding the model to integrate seamlessly with additional social media platforms and digital content creation tools is another priority. This will involve developing plugins or APIs that can be easily adopted by various platforms, broadening the reach and applicability of the Storify model.

Final Note: The ongoing evolution of the Storify model promises to not only enhance the music recommendation system but also contribute to the broader field of multimedia content analysis and user engagement strategies. With continued research and development, Storify aims to set new standards in the integration of AI with digital media experiences.

ACKNOWLEDGEMENTS

We extend our deepest gratitude to Dr. Rajiv Ratn Shah, Mr. Adarsh Pandey, and Mr. Kapuriya Sureshbhai for their invaluable guidance and constant support throughout the development of the Storify model.

Their willingness to share their vast knowledge and experience has been a significant catalyst in not only advancing this project but also in fostering a learning environment that encourages innovation and critical thinking. We are sincerely thankful for their patience, motivation, and enthusiastic engagement throughout this journey.

Additionally, we wish to express our appreciation to our peers and colleagues at the institute who provided feedback and participated in the testing phases of our system. Their contributions have been immensely helpful in refining the Storify model to better meet user needs.

This project would not have reached its fruition without such collaborative efforts, and we are profoundly thankful for every bit of expert advice and moral support received during this research.

10 GITHUB LINK

- Code

11 REFERENCES

11.1 Datasets

- FER-2013 Dataset: https://www.kaggle.com/datasets/msambare/fer2013
- Human Action Recognition (HAR) Dataset: https://www.kaggle.com/datasets/meetnagadia/human-action-recognition-har-dataset
- Pipleline Diagram: https://lucid.app/lucidchart/bbccba72-b6ad-412d-b9c3-0630a33386ee/edit?viewport_loc=78%2C132%2C1675%2C851%2C0_0&invitationId=inv_0f0b3945-6cc0-4f71-a23f-afed3ca38dd0

11.2 Publications

[1] Facial Emotion Recognition:State of the Art Performance on FER2013 Facial Emotion Recognition: State of the Art Performance on FER2013

[2] Social Media Advertising Market Insights - A comprehensive report on the social media advertising market trends and forecasts.

[3] Schedl, M. (Year). Deep Learning in Music Recommendation Systems. *Journal Name*, *Volume*(Issue), Pages. DOI

[4] Song, Y., Dixon, S., & Pearce, M. (Year). A Survey of Music Recommendation Systems and Future Perspectives. In *Proceedings of the International Symposium on Computer Music Modelling and Retrieval*.

[5] Schedl, M. (Year). Integrating Music Content, Music Context, and User Context for Improved Music Retrieval and Recommendation. In *Proceedings of the MoMM2013 Conference*.

[6] APA: Culture and the categorization of emotions - Russell, J. A. (1991). Psychological Bulletin, 110(3), 426–450.

[7] https://ieeexplore.ieee.org/document/9847888

[8] https://journals.sagepub.com/doi/full/10.1177/2041669520961123

[9] https://medium.com/@leea34570/impact-of-music-media-in-social-media-marketing-ed19dc268413 - Medium: Impact of Music Media in Social Media Marketing

[10] https://www.mdpi.com/1660-4601/19/24/16637