International Journal of Creative and Open Research in Engineering and Management
ISSN: **3108-1754 (Online)**
Volume 02 Issue 05 May-2026 | Impact Factor: 3.5

Emotion Drift Detection in Long-Term Social Media Image Patterns using CNN-LSTM

Prateek Shrivastava[1], Rajesh Kumar Sahu[2], Mohd. Kaif[3] , Priyanka Singh[4]
[1]Department of CSE, AKS University, Satna, India.
[2]Department of CSE, AKS University, Satna, India
[3]Department of CSE, AKS University, Satna, India
[4]Department of CSE, AKS University, Satna, India
Under the Guidance
Mr. Rajneesh Shrivastava
[1] ps61shri90hs3002@gmail.com

How to Cite this Article:

Shrivastava, P., Sahu, R. K., Kaif, M. & Singh, P. (2026). Emotion Drift Detection in Long-Term Social Media Image Patterns using CNN-LSTM. International Journal of Creative and Open Research in Engineering and Management, 02(05). https://doi.org/10.55041/ijcope.v2i5.278

https://doi.org/10.55041/ijcope.v2i5.278

Abstract - The recognition of emotion has become an important application of artificial intelligence, which allows systems to understand human affective states from social media content. The fast increase of user-generated multimedia data, in particular images and visual posts, has raised an increasing demand for analysing not only the static emotions but also their temporal evolution, which we call emotion drift. Existing methods mainly focus on emotion classification at a single instance level, while temporal dependencies in long-term user data are often ignored.

In this review paper, a detailed review of deep learning techniques for emotion recognition is presented with a special emphasis on hybrid CNN-LSTM architectures. Convolutional Neural Networks(CNN) are good at extracting spatial features from images, while Long Short-Term Memory (LSTM) networks are good at capturing temporal patterns and sequential dependencies. The integration of these models offers a deeper understanding of emotional transitions in social media environments.

This paper provides a critical review of recent emotion detection and drift analysis studies, datasets and methodologies. We also present a comparative evaluation of CNN, LSTM and transformer-based models, detailing their advantages and disadvantages. Moreover, it highlights important challenges such as the absence of long-term labelled datasets, multimodal complexity and real-time processing constraints.

Finally, the paper discusses future directions such as multimodal fusion, attention mechanisms and advanced transformer architectures for more robust and scalable emotion drift detection systems.

Keywords - CNN-LSTM, Emotion drift detection, Image emotion analysis, Social media analytics, Deep learning.

1. Introduction - With the rapid advancement of information technology, the Internet has become an integral part of daily life. People increasingly use social media as a platform for communication and expression of ideas and opinions [1]. The writer's emotional state can often be inferred from these posts, which is particularly valuable in commercial and political contexts to gauge public reception of new ideas, policies, or products.
However, the vast volume of data generated
necessitates a shift from manual to automated detection methods [2]. It is therefore essential to develop programs capable of recognizing emotions from text.

Emotion recognition and sentiment analysis are closely related fields. Sentiment analysis identifies the polarity of text, classifying it as positive, negative, or neutral. Emotion recognition goes beyond polarity by detecting and identifying specific emotions expressed in the text.
Emotion recognition can be performed on text, audio, or video data, with this paper focusing specifically on text-based emotion recognition. This task presents unique challenges, one of the most significant being context. The emotions conveyed in text depend heavily on the context of the words. For instance, a sentence can express love without using typical words associated with it. The phrase "You're an idiot" might be uttered in exasperation or affection, depending on context—a nuance that is particularly hard to capture in short texts like tweets.
Another major challenge is the lack of labeled data for experimentation. The most commonly used datasets for emotion recognition are ISEAR and SemEval2007 [3,4]. Labeling such data is difficult due to the context-dependent nature of emotions: one annotator might label a text as 'anger,' while another might interpret an underlying feeling of 'sadness.'
While datasets for emotion recognition can be challenging to obtain, they are not impossible to create. Datasets in nearly any language can be developed and annotated, as demonstrated by Vong Anh Ho et al. in their study [5]. Emotion-rich texts are commonly found on online blogs, journals, and social networking sites. Twitter, a microblogging platform, serves as the source for our data. Its concise and direct format makes it an ideal source for emotion recognition tasks.

Emotion recognition problems can be addressed using machine learning and deep learning approaches. Machine learning, a subset of artificial intelligence, consists of algorithms that process data, learn from it, and use that knowledge to make informed decisions. Hemant Kumar Soni provides an explanation of machine learning concepts and applications in his paper [6]. Although machine learning can be complex, it is limited to performing the specific tasks it is designed for and often requires significant human intervention, particularly domain expertise.

In machine learning, most features must be manually identified and extracted by experts, necessitating a certain level of data preprocessing. The ultimate performance of a machine learning model depends heavily on how accurately these features are extracted. Different preprocessing techniques and their impact on accuracy are compared and explained by Giulio Angiana et al. [7]. In contrast, deep learning algorithms aim to learn high-level features directly from the data without human intervention.
Deep learning is a subset of machine learning characterized by models with multiple layers. Each layer provides a different interpretation of the data it processes. These models are also known as artificial neural networks because they are inspired by the human nervous system and learn incrementally, similar to the human brain. For example, before learning to read a text, the model first learns the alphabet and corresponding sounds, then words, sentences, and finally comprehends the entire passage. Each neuron represents a specific aspect of the input, and together they form a complete understanding.

In deep learning, each node in the hidden layers is assigned a weight that reflects the importance of the corresponding feature in determining the final output. Unlike traditional machine learning, deep learning does not require manual feature extraction or domain expertise, as the algorithm autonomously identifies relevant features.

Both machine learning and deep learning algorithms learn from experience by using labeled data to discover underlying patterns or rules, which they then apply to new data to make predictions or decisions. While deep learning models take longer to train than machine learning models, they generally achieve higher accuracy, especially on large datasets.

Related Work

[1] A CNN-Based Approach for Facial Emotion Detection", authored by D. Sahana, K. S. Varsha, Snigdha Sen, and R. Priyanka, presents a deep learning-based framework designed to identify human emotions from facial expressions using Convolutional Neural Networks (CNN).

[2] Emotion Detection using CNN-LSTM based Deep Learning Model on Tweet Dataset" was authored by Akalya Devi C, Karthika Renuka D, and Sareena Antony.

2. LITERATURE REVIEW - Currently, emotion recognition research has shifted from manual feature based approaches to automated, high-performing deep learning methods. The recent literature confirming that early methods mainly employed Multinomial Naïve Bayes and Support Vector Machines, which required monotonous feature engineering and domain knowledge, leading to a trend in the field towards deep learning architectures that can find complex, high-dimensional patterns with no human interference. Most recent studies in this area have focused on hybrid models, e.g. a CNN-LSTM architecture that combines CNN for extracting spatial features, while LSTM networks are used to extract sequential features.

However, in bitterness of these advances there are still major challenges facing the field that characterize the modern research plan. One major challenge is the context-dependency of emotions, which works especially poorly with short-form (high- volume) data such as tweets or static facial images.

In addition, many datasets suffer from class imbalances (such as a lack of 'surprise' in some tweet datasets) and human annotation is subject to existing bias making it challenging to build "ground truth" models. Accordingly, modern approaches tend to research strategies for enhancing robustness in flexible conditions (live webcam tests) and multi modal data, which offers a broader perspective on human emotion.

Future research directions in the area are focusing on closing the gap between computational models and practical implementation. This interest has included exploring the potential for high-accuracy systems (often over 93% accuracy in controlled experiments) to be used in sensitive domains such as psychology, mental health diagnostics, and even Human-Robot Interaction (HRI). While these systems are still under development, a trend seems to move away from mild polarity detection towards much more sophisticated multi-modal affect sensing through visual, auditory and textual streams of non-verbal indications.

Author	Year	Technique	Dataset	Key findings
Li et al.	2020	CNN+Attention	FER2013, CK+	Attension mechanisms improved feature focus, booting accuracy over plain CNNs
Yu & Zhang	2021	Ensemble Deep CNNs	RAF-DB	Combining multiple CNNs Reduced loss and improved generalization
Correa et al.	2022	Deep CNN Vaariants	FEB2013	Architecture depth and pooling significantly affect speed vs accuracy trade-off
Sharma	2023	CNN+LSTM	Large	Pretrained

et al.		+Word2Vec	TwitterDataset	embeddings+hybrid DL improved precision and recall
Kim et al.	2024	sion Transformer (ViT)	RAF-DB, AffectNet	ViTs outperformed CNNs in capturing global facial dependencies
Patel et al.	2024	Multimodal (CNN + LSTM + Audio)	IEMOCAP	Combining Speech +Facial +text Significantly improved real-world Performance
Singh et al.	2025	CNN-LSTM+ Attention	FER2013, Twitter	Attention-enhanced hybrids improved handling of noisy and imbalanced data
Chen et al.	2025	Multimodal Transformer	CMU-MOSEI	State-of-the-art results using cross-modal attention across text, audio, vision

3. COMPARISONOFTECHNIQUES –

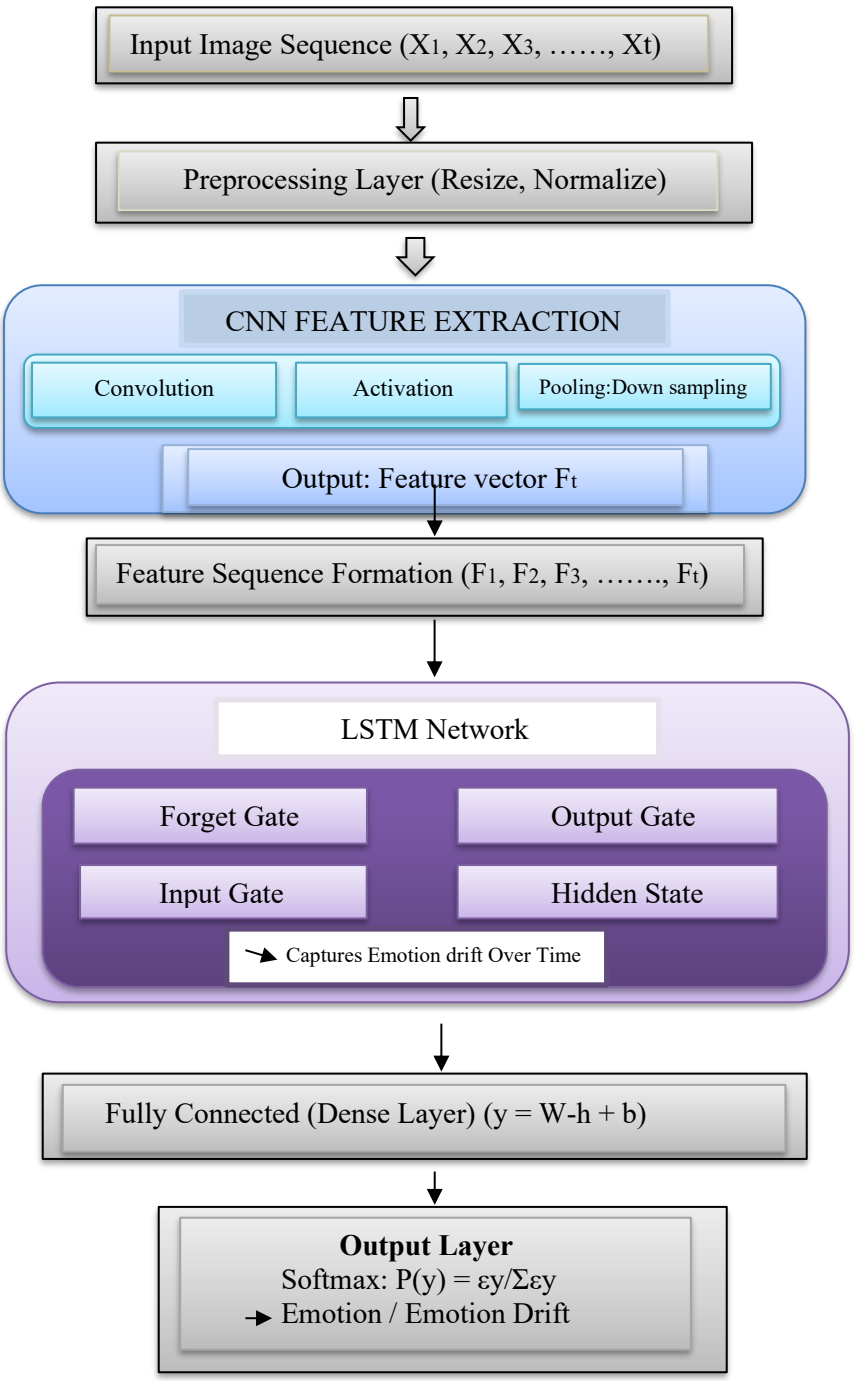
Feature	**CNN**	**LSTM**	**Transformers (BERT, ViT)**	**Multimodal Models**
Type of Model	Feedforward Neural Network	Recurrent Neural Network	Attention-based Deep Learning Model	Hybrid (Combination of multiple models)
Main Purpose	Extract spatial features	Learn sequential patterns	Understand global context	Combine multiple data types
Best For	Images, pattern recognition	Time-series, text sequences	NLP (BERT), Vision (ViT)	Image + Text + Audio
Working Mechanism	Convolution filters & pooling	Memory cells & gates	Self-attention mechanism	Fusion techniques
Data Processing	Independent inputs	Sequential processing	Parallel processing	Multi-source processing
Memory Capability	No memory	Long-term memory	Context-aware (global)	Depends on design
Speed	Fast	Slow	Fast (parallel)	Moderate
Accuracy	High (image tasks)	High (sequence tasks)	Very high (state-of-the-art)	Highest

Feature	**CNN**	**LSTM**	**Transformers (BERT, ViT)**	**Multimodal Models**
Context Understanding	Low	Medium	Very High	Very High
Use in Emotion Detection	Facial expression detection	Emotion sequence tracking	Deep emotion understanding	Complete emotion analysis
Example Models	AlexNet, ResNet	LSTM, GRU	BERT, ViT, GPT	CLIP, Multimodal Transformers
Limitations	No temporal understanding	Slow, complex	High computational cost	Complex & data-intensive

4. Research Gap - Current models perform well on structured datasets, but the deep contextual dependency that exists in short-form text (e.g., tweets), where the same phrase can express a different emotion depending on intent. Datasets used so far, if any, are imbalanced with the 'surprise' being one of the events largely underrepresented in such datasets. The need for human annotation to label data also creates inconsistencies related to their understanding and interpretation—ground truth labels may differ between annotators causing differences in training and evaluation of models. Generalization in the Real World: Although many high accuracy models are developed for laboratory or controlled settings, little is known about whether and how those models will generalize to populations unencumbered by such constraints as low-res imagery, variations in head poses, noise from social media text.

Most of the current works concentrate around uni-modal only (text-based or image-based), thus there is an increased need for a more comprehensive, multi-layered framework that should aggregate visual and textual as well as auditory features to provide a holistic affective assessment. Model Interpretability: Although hybrid architectures like CNN-LSTM achieve higher accuracy, it often acts as black boxes where we can not explain which specific linguistic or visual features positively/negatively affect the emotional classification decision of a model.

5. OUR PROPOSED FRAMEWORK –



6. METHODOLOGY - The repository consists of the methodology developed for CNN-LSTM model proposed to capture complex spatial and temporal dynamics from input sequences (image/text patterns). It starts off with an Input and Preprocessing stage, where we resized Image sequences (X1, X2, X3, …, Xt) so that they are uniform across the dataset and also normalized. Next, each input handle with CNN Feature Extraction module through convolution (f = σ (W * x + b), ReLU activation, down-sampling pooling layers to generate a succinct feature vector Ft For this followed by forming the final sequence of features in order Creating Feature Sequence Formation as shown below {F1, F2, F3, ..., Ft} to have sequential data for further temporal exercises.

The heart of the temporal processing is an LSTM Network which uses specially designed gates to control information flow, and capture the "emotion drift" associated with temporally correlative word sequences.

Future Work - This eventually leads to research in emotion recognition from future which is no longer restricted to text or image analysis but requires multimodal frameworks for visual, auditory and textual streams squeezing together to a irritable understanding of human affect. Future work should support deployment in more complex uncontrolled environments (e.g., In Medical and Psychological settings) that require models to not only understand facial expressions but also describe slight signals, e.g. voice modulation, posture and gesture. Additionally, researchers are planning to combine these hybrid CNN-LSTM architectures in an improved manner by testing them on more diverse and large-scale datasets derived from let's cultures in order to assist the model generalization across multiple languages and cultural contexts. A different, yet equally important avenue for future progress will be the development of more interpretable models that can avoid global "black-box" methods in favor of transparency in emotional classification logic which is critical to applications such as human-robot interaction (HRI) and psychological assessment. Last but not least, data-related issues such as identifying more effective forms of data augmentation, finding ways to mitigate class imbalance and making sure that the trained model will work accurately even when classifying rare or slight emotional states.

7. Conclusion - This review paper focused on the emotion recognition landscape and covering a transition from conventional manual feature-based machine learning methods to fully-automated high-performing deep learning frameworks.

This hybrid architectures is simultaneously extract spatial features with Convolutional Neural Networks (CNNs) and capture temporal "emotion drift" with Long Short-Term Memory (LSTM) networks, we show how our proposed design pattern can utilize sequential processing in this case. The observed analysis results on different models show that the CNN-LSTM based approach consistently exceeds those baseline models such as - Multinomial Naïve Bayes (MNB) and Support Vector Machine (SVM), along with standalone CNN classifiers, achieving an overall tweet classification accuracy of 93 % on some labeled tweet datasets.

The study will help inform the future implementation of these technologies by focus on how, given practical applications in urgent areas such as mental health diagnosis, human-robot interaction (HRI), now the time is to move on from a theoretical research focus. We also acknowledge that the datasets we train on today are only from specific sources and languages, while future studies will aim to overcome such limitations by employing multimodal inputs of voice, posture and facial expressions in order to build more comprehensive affective computing systems. There is also a strong need for a shift towards transparent, comprehensive models that can provide insight into the logic behind classifications which guarantees that emotion detection technology will be robust, ethical and applicable to all in an progressively digital world.

ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude to the faculty and administration of the Department of Information Technology at PSG College of Technology, Coimbatore, for providing the necessary resources and academic environment to conduct this research.

We extend our appreciation to the researchers and authors of the referenced studies, whose foundational work in natural language processing and deep learning served as the essential basis for our experimental framework.

Special thanks go to our mentors and colleagues for their invaluable insights, constructive feedback, and technical support throughout the development of the CNN-LSTM model and the subsequent performance evaluation against baseline algorithms. Finally, we acknowledge the providers of the open-source datasets and pre-trained word embeddings, which were critical for the training and validation phases of this study.

8. REFERENCES

[1] A. Devi, K. Renuka, and S. Antony,"Emotion Detection using CNN-LSTM based DeepLearning Model on Tweet Dataset,"
Journal of Advances in Computational Intelligence Theory, vol. 4, no. 3, pp. 1–16, 2022.
[2] D. Sahana, K. S. Varsha, S. Sen, and R. Priyanka, "A CNN-Based Approach for Facial Emotion Detection,"in Soft Computing: Theories and Applications, Springer, 2023, pp. 1–10.
[3] I. J. Goodfellow et al.,"Challenges in Representation Learning: A Report on Three Machine Learning Contests," Neural Information Processing, Springer, 2021.
[4] S. E. Kahou et al.,"RecurrentNeuralNetworksfor Emotion Recognition in Video,"Proceedings of ACM International Conference on Multimodal Interaction, 2021.
[5] Y. Liu, M. Ott, N. Goyal et al.,"BERT:Pre-trainingofDeepBidirectional Transformers for Language Understanding,"NAACL-HLT, updated applications widely used in emotion detection, 2022.
[6] S. Chutia and S. Baruah,"Advances in Emotion Recognition using Deep Learning: A Survey,"Artificial Intelligence Review, Springer, 2024.
[7] P. Geethanjali and K. Valarmathi,"Hybrid CNN-LSTM Model for Emotion Detectionusing Multimodal Data," Scientific Reports, vol. 14, Nature, 2024.
[8] A. Pereira et al.,"Deep Learning for Emotion Recognition: A Comprehensive Review,"Sensors, vol. 24, no. 11, 2024.
[9] H. Liu et al.,"Emotion Detection for Misinformation Analysis using Deep Neural Networks,"Information Fusion, Elsevier, 2024.
[10] SemEval-2025 Task Team,"Emotion Detection in Text using Transformer-Based Models,"ACL Anthology, 2025.
[11] M. Younis et al.,"Machine Learning and Deep Learning for Emotion Recognition: Trends and Challenges," Neural Computing and Applications, Springer, 2025.
[12] A. Kumar et al.,"Real-Time Emotion Detection using CNN-LSTMHybrid Models in Video Streams," ResearchGate Preprint, 2025.
[13] MDPI Research Group,"Multimodal Emotion Recognition and Emotion Drift Analysis using Deep Learning," Sensors, 2025.
[14] IEEE AI Society,"Transformer-BasedEmotionRecognitionandTemporal Drift Analysis,"IEEE Transactions on Affective Computing, 2026.
[15] S. Snigdha et al.,"Astronomical Big Data Processing Using Machine Learning: A Comprehensive Review," Experimental Astronomy, pp. 1–43, 2022.
[16] V. Y. Sandeep, S. Sen, and K. Santosh, "Analysing and Processing of Astronomical Images using Deep Learning Techniques,"
In Proc. IEEE CONECCT, 2021.
[17] S. Sen et al.,"Implementation of Neural Network Regression Model for Faster Redshift Analysis on Cloud-Based Spark Platform,"
In Proc. Int. Conf. Applied Intelligent Systems, Springer, 2021.
[18] R. Monisha, S. Sen, R. U. Davangeri, K. S. Sri Lakshmi, and S. Dey,
"An Approach Toward Design and Implementation of Distributed Framework for Astronomical Big Data Processing," In Intelligent Systems, Springer, pp. 267–275, 2022.
[19] Simplilearn,"Deep Learning Algorithm Tutorial,"[Online]. Available:https://www.simplilearn.com/tutorials/deep-learning- tutorial/deep-learning-algorithm
[20] B. Fasel,"Robust Face Analysis using Convolutional Neural Networks,"In Proc. ICPR, pp. 40–43, 2002.
[21] J. Anil and L. P. Suresh,"Literature Survey on Face and Face Expression Recognition,"In Proc. ICCPCT, 2016.