Cureus Journal of Computer Science

Open Access Original Article

A Machine Learning-Based Approach to Personification and Safety Measures in Social Media Applications

Received 06/11/2025
Review began 06/27/2025
Review ended 08/12/2025
Published 10/08/2025

DOI:
https://doi.org/10.7759/s44389-025-07434-4

Prema Sahane [1] , Sonali Rangdale [2] , Sowmyashree H. Srinivasaiah [3] , Deepali Patil [4] , Sonali Patil [4] , Uttam U. Deshpande [5]

1. Department of Computer Engineering, JSPM's Rajarshi Shahu College of Engineering, Pune, IND 2. Department of Information Technology, G H Raisoni International Skill Tech University, Pune, School of Engineering and Technology, Pune, IND 3. Department of Computer Applications, Bangalore Institute of Technology, Bengaluru, IND 4. Department of Computer Engineering, Dr. D. Y. Patil School of Science and Technology, Pune, IND 5. Department of Electronics and Communication Engineering, KLS Gogte Institute of Technology, Belagavi, IND

Corresponding author: Uttam U. Deshpande, uttamudeshpande@gmail.com

Abstract

The amount of time an average person spends on social media platforms for collecting, sharing, and storing data is exponentially increasing. Social media's rapid growth has brought up serious problems of user impersonation, identity theft, and insecure interactions. As a result, the people's social media data privacy and safety management concerns are drawing urgent attention. Machine learning algorithms' ability to identify customers' usage patterns and their content enables them to identify cyberbullying and fake news. This research provides a multimodal machine learning-based framework for safeguarding user interaction through trustworthy persona detection and safety classification. To identify harmful profiles and misleading material, we utilize diverse text, images, videos, and behavioral pattern data to test the state-of-the-art learning Convolutional Neural Network, Random Forest, Support Vector Machine, Logistic Regression, and Adaptive Boosting algorithms. Among all, Convolutional Neural Network surpassed the rest with an accuracy of 83.1% on benchmark datasets like Face Detection Data Set and Benchmark (FDDB), Annotated Facial Landmarks in the Wild (AFLW), and the Deepfake Detection Challenge dataset. Our approach allows real-time user emotion and activity recognition, providing a scalable and adaptable approach to enhance safety on social media platforms. This work bridges significant gaps in authenticating content genuineness, trust, and variations in digital communication environments.

Categories: Security Governance and Compliance, Machine Learning (ML), Explainable AI
Keywords: multimodal machine learning, online safety, social media security, persona identification, cnn, fddb, aflw, deepfake detection

Introduction

Over the past few years, the world has observed widespread social media usage that has completely transformed the way individuals communicate, connect, share, and interact online. Social media platforms such as Facebook, Twitter, Instagram, or WhatsApp have allowed various users to connect at an extraordinary scale. Despite all the benefits that these social media applications offer, they have introduced significant user safety and privacy concerns. "Personification," which can assign or imitate human-like traits or identities through algorithms, user profiles, or platform behaviors, is another major growing concern. Personification, as defined in this study, refers to identity falsification techniques in which automated systems or human users exhibit false or deceptive identities. Posing as someone else, making fake profiles, or mimicking behaviour patterns to seem genuine are a few examples. This interpretation is not the same as algorithmic personalization, which focuses on adjusting platform recommendations or content to a user's actual preferences and past interactions. Instead of improving the distribution of personalized content, this initiative focuses on identifying and combating false identification practices.

The safety of users, platform integrity, and trust are all significantly impacted by this phenomenon. New developments in machine learning offer [1] encouraging possibilities for identifying inconsistency in behaviour [2], examining user interaction trends, and protecting digital identities in real-time [3]. The integration of multimodal data - text, video, voice, and behavioral metadata - for safe user profiling has not received much attention in previous research, even though many studies have concentrated on privacy, sentiment classification, or fake news identification.

Despite sufficient progress made by the research community [4,5] on user trust, privacy issues, ethical considerations, and the psychological impacts of social media topics, there is still a lack of comprehensive machine learning based methods that integrate these topics to provide security against user privacy and personification risks [6]. The proposed study offers a systematic literature review revealing the use of

important, ignored topics that can be integrated with current approaches. The study provides a strong machine learning framework foundation for creating strong and safe persona detection solutions. The main contributions are summarized below.

i. Propose a secure, AI-driven framework that integrates Convolutional Neural Network (CNN), Random Forest, Support Vector Machine (SVM), AdaBoost, and Logistic Regression algorithms for classifying social media personas.

ii. Combine behavioral, emotional, and media content-based characteristics to evaluate multiple machine learning models (CNN, Random Forest, SVM, AdaBoost, Logistic Regression) performances.

iii. Validate these models on real-world video-based persona identification on facial landmark recognition, Face Detection Data Set and Benchmark (FDDB), Annotated Facial Landmarks in the Wild (AFLW), and Deepfake Detection Challenge benchmark datasets.

Literature survey

Machine Learning Models for Safety in Social Media Applications

The popularity of social media applications has increased online communication in recent years. As a result of the growing amount of private information published on these platforms, protecting user privacy and safety is necessary. Personification of social media users using machine learning techniques is one way to achieve safety in social media applications. Machine learning techniques for persona classification and identification are used for overcoming possible risks like cyberbullying, harassment, or false information by utilizing sophisticated algorithms and data analysis approaches [1,2]. In this research, the advantages and difficulties of applying machine learning techniques for personification and safety in social media are examined [3].

Understanding Personalization and Safety in Social Media Platforms and Its Importance

Social media platforms are crucial in connections within society in the current digital era. As data shared on applications like Facebook, Instagram, and Twitter platforms increases, protecting user privacy and safety is more important. Users' experiences can be improved if they feel safer communicating over social media. It is important to make sure safety precautions are in place to minimize users from privacy violations [1,4-6]. Understanding the significance of safety and personalization in social media platforms allows machine learning techniques to be applied to efficiently customize information to a person and protect user data and achieve proper utilization of social media applications [7].

Overview of Machine Learning Techniques Used for Personification in Social Media

To improve user experience and safety in social media, personification using machine learning techniques is more important. Methods like Linear Regression and SVM perform the classification of user interactions, persona, and behaviour to predict recommendations and classify content. Social media platforms like Instagram, Facebook, etc. can increase user engagement and satisfaction by using machine learning models for feature selection, extraction, classification of behaviour, analysis of sentiments, natural language processing, audio and video classification, and to predict content for specific users [3,7,8]. Machine learning algorithms can also be used to identify security threats. Hence, these methods are essential for enhancing user experience and maintaining safety in the online environment [3,9].

Ensuring User Safety Through Machine Learning Algorithms on Social Media Platforms

To keep safe online communication in social media applications requires social media companies to integrate machine learning algorithms to ensure users' safety. Social media companies can identify and stop bad behaviors by using a specific learning model. Real-time analysis and classification of persona and patterns can spot negative class behaviour [7,10,11]. Additionally, by suggesting relevant content and removing objectionable content, machine learning techniques can be utilized for better user experiences. With learning methods, social media platforms give personification for users and make online communication safer [11,12].

Machine learning methods are proposed for safety after persona classification in social media applications like Facebook and WhatsApp. Previous studies suggested a Modified Inception Resnet V4 CNN (MInReCNN) for personalized recommendations [13], while one article focuses on women's safety, using machine learning classification techniques like SVM, Naive Bayes, and AdaBoost for analysis of sentiments [14]. The author has given the issues of fake profiles, proposing methods from machine learning algorithms to detect and identify them [15]. To detect misleading identities on social media platforms, features from profile images and gender-derived features are studied and reviewed using machine learning techniques for personality identification [16,17]. The author suggested a deep

multimodal fusion approach for user profiling and developed a multimodal deep framework for identifying critical social media posts [18,19]. These studies collectively highlight the possibility of machine learning in addressing various safety and personification challenges in social media.

Future Prospects and Encounters in Using Machine Learning for Personification and Safety in Social Media

Machine learning algorithms have a lot of potential to comprehend human behaviour and preferences better as technology develops, which could result in more individualized interactions. This could improve social media platform user experience and engagement. But some issues need to be resolved, like protecting privacy, preventing algorithmic bias, and preserving openness in the decision-making processes [3,20,21]. Furthermore, the rapidly changing nature of social media applications and the massive amount of data being produced pose persistent hurdles for machine learning techniques to adapt and successfully handle personification and safety-related issues [12].

Design issues

Learning a database that includes media data like text messages, images, and video frames is crucial because these types of content are commonly used by social media apps like WhatsApp, Facebook, and Twitter. The goal is to develop a data analytics system that can improve success rates by analyzing patterns and classifying content effectively. However, the training process for this type of data analytics can be inefficient due to the lengthy training time required. To address this issue, meta-heuristic algorithms are used to find optimal solutions at a reasonable computing cost. Some meta-heuristic algorithms can be reviewed for their use in direct, indirect, and hidden data. For example, identifying patterns in text crowding can be challenging. The decision to move outside or remain indoors for health reasons is an essential part of the dataset. Different objects are arranged in a specific way to create rules and patterns that govern the placement of content.

Materials And Methods

Methodology

We propose a four-stage pipeline containing data preprocessing, multimodal feature fusion, model training, and post-classification evaluation for developing a robust social media user personification and security classifier.

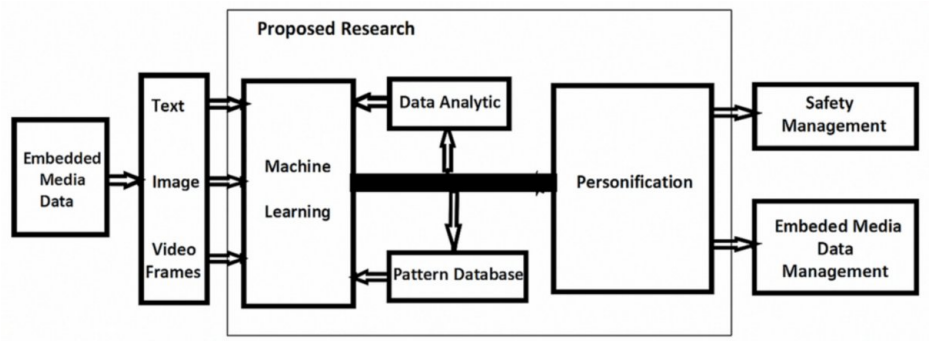

FIGURE 1: Proposed methodology

As shown in Figure *1* of the proposed research methodology, the embedded or social media application contents, such as images, texts, video frames, and audio, are taken as the input focus. Social media applications like Twitter, Instagram, WhatsApp, Facebook, etc., use the same media content. By analyzing a case study of Facebook, it is possible to incorporate features such as gender, age, and behaviour into the object's sub-feature. This can help achieve personification and safety. To create a feature vector that achieves this, input data is given to various algorithms like K-means, DBSCAN, and principal component analysis. From the results, it is observed that K-means performed well in identifying the top terms of cluster centers, but it did not correlate well with reviews in some instances [9]. On the other hand, DBSCAN grouped all behaviors into one cluster when analyzing the following 11 distinct feature variables, resulting in poor performance.

i. Valence - a numeric value indicating the pleasantness of the detected emotion.

ii. Arousal - a score reflecting the emotional activation or intensity.

iii. Dominance - a measure of the perceived control or influence within the emotional state.

iv. Facial landmarks - coordinate points representing key regions of the face, such as eyes, nose, and mouth, used for geometry-based analysis.

v. Facial action units - codes describing specific facial muscle movements linked to expressions.

vi. Sentiment polarity - categorical or numeric representation of text sentiment as positive, negative, or neutral.

vii. Object detection flags - binary indicators marking the presence of predefined objects in images or frames.

viii. Pitch - the fundamental frequency derived from speech segments.

ix. Energy - the loudness or amplitude level of the audio signal.

x. Behavioral patterns - metrics describing frequency, type, and timing of user interactions or activities.

xi. Demographic attributes - non-identifiable metadata such as age group and gender category.

Hierarchical clustering, when used with BOW and TFIDF, failed to identify clusters and divide them unevenly. However, when using average word vectors, clusters were evenly divided and identified. Despite these challenges to determine the user based on hierarchical formation is still hard. To evaluate and contrast the effectiveness of the clustering methods used, specifically K-means, DBSCAN, and hierarchical clustering, we used the silhouette score and cluster purity as measurements. While cluster purity represents the percentage of samples in a cluster that are members of the same reference class, the silhouette score shows how correctly each data point falls into its allocated cluster in relation to nearby clusters. By combining these measures with visual assessment of cluster assignments, it was possible to determine when clusters were significant and well-separated, as well as when overlaps between behavioral and emotional traits degraded the grouping's overall quality.

Construction of Two-Dimensional Valence-Arousal Emotion Map

The spatial distribution of emotional states was represented using a two-dimensional valence-arousal framework. According to this method, emotions are represented as coordinates inside a plane, with the vertical axis (arousal) spanning from low to high activation levels and the horizontal axis (valence) ranging from unpleasant to pleasant affect. Both the intensity and the positivity or negativity of emotions can be visualized, thanks to this dimensional mapping.

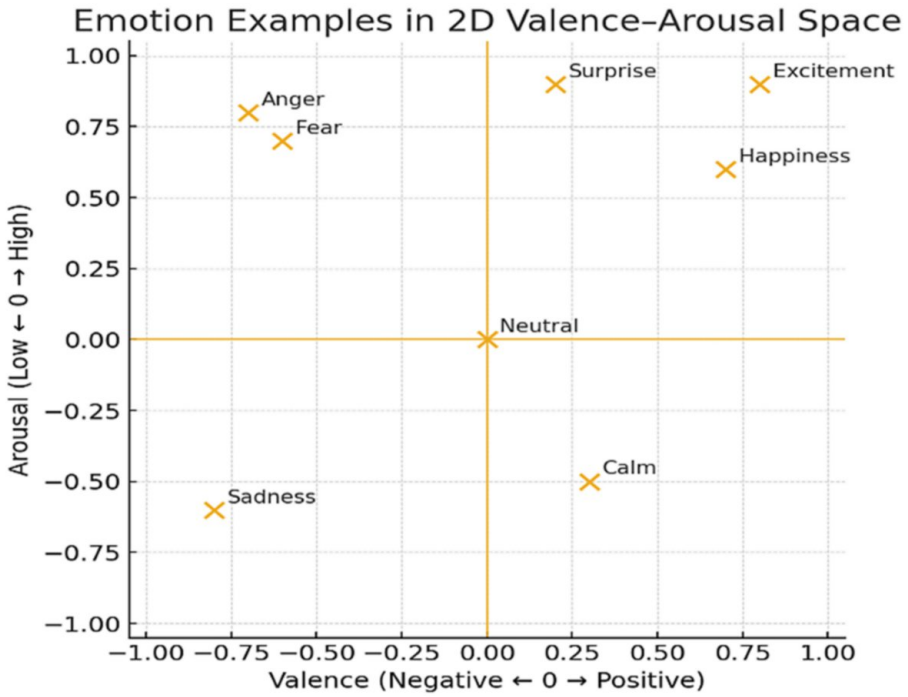

FIGURE 2: Valence–arousal map showing emotions by pleasantness and activation, with examples in each quadrant and neutral near the origin

Figure 2 shows the plotting of the identified emotions based on their corresponding locations in this space. Anger, for example, is found in the negative high-arousal quadrant, serenity is found in the positive low-arousal quadrant, excitement is found in the positive high-arousal quadrant, and melancholy is found in the negative low-arousal quadrant. Neutral states, which represent balanced valence and moderate arousal, are located close to the origin. Mathematical equations are used to identify and detect emotions from videos. For example, in Equation (1), $V(t)$ describes how the valence (mood swings or sentiment anomalies) of an emotion (positive/negative) changes over time 't':

$$\boldsymbol{V}(t) = \boldsymbol{V}_{\text{init}} + (\boldsymbol{ROC} \cdot t) \qquad (1)$$

where $V_{(\text{init})}$ represents the baseline (Initial valence) emotional state, and emotional volatility (valence shifts over time) is obtained from Rate of Change (ROC). The time or frame index 't' models emotions from video or time sequence data.

Multimodal Feature Fusion

Physiological markers, which can be used to categorize and predict emotional states, are the basis of another method for identifying emotions. Variables like skin conductance, heart rate, and user activity levels are all included in these models. Three primary categories of user data are also taken into consideration:

• Emotional features are obtained from speech tone or prosody patterns, sentiment scores from text captions, and facial expression analysis (AFLW dataset).

• Behavioral features, such as the frequency of content sharing, likes, scrolling habits, and login history, are taken from the metadata of user interactions.

• Media features include visual descriptors, facial embeddings, and objects identified by CNN-based tagging.

These feature sets are integrated into a single combined vector, with the process mathematically represented in Equations (2) and (3).

Feature Vector X_{comb}: Here, X_{comb} represents a combined feature vector that can be obtained by

concatenating the emotional features X_{emo}, media features X_{media}, and behavioral features X_{behave} for representing individual users' personas.

$$\boldsymbol{X}_{\text{comb}} = [\boldsymbol{X}_{\text{emo}} \,\|\, \boldsymbol{X}_{\text{media}} \,\|\, \boldsymbol{X}_{\text{behave}}] \qquad (2)$$

Standardizing $X_{normalize}$: For the dataset to write similar scale for all features, we can normalize the total feature vector X_{comb} represented as

$$\boldsymbol{X}_{\text{normalize}} = \frac{\boldsymbol{X}_{\text{comb}} - \boldsymbol{\mu}_{\boldsymbol{X}_{\text{comb}}}}{\boldsymbol{\sigma}_{\boldsymbol{X}_{\text{comb}}}} \qquad (3)$$

where $\mu_{X_{comb}}$ describes the mean value of each feature across the dataset, and σX_{comb} representing the standard deviation of each feature.

Classification

To train CNN, SVM, Random Forest, AdaBoost, and Logistic Regression ML models, we used a normalized vector $\hat{Y}$ to predict a label (Real/Fake) or classify (Risky/Not Risky) as represented using Equation 4.

$$\hat{\boldsymbol{Y}} = f_{\text{classifier}}\left(\boldsymbol{X}_{\text{normalize}}\right) \qquad (4)$$

Here, $f_{classifer}$ represent the trained classifier function of the above-mentioned ML models on a $X_{normalize}$ input features obtained from Equation (3). The best-performing classifier (CNN) is further optimized using grid search on hyperparameters: learning rate, kernel size, dropout rate, and batch size.

Testing Model: To test the performance of the model $\text{Model}_{\text{T}}\text{est}(f)$, evaluation methods like mean squared error (MSE), accuracy, Recall, Precision, etc., are used on unseen test data. After evaluation, we have refined the model by iterating on feature selection and model parameters to improve its performance, as shown in Equation (5). Representation of this in an iterative process is

$$\text{Model\_Test}(\boldsymbol{f}) = \text{Eval}\left(\boldsymbol{f}\left(\boldsymbol{X}_{\text{input\_test}}\right)\right) \Rightarrow \text{Refine}\left(\boldsymbol{f}, \boldsymbol{\theta}\right) \qquad (5)$$

The input test data is represented as $X_{input\ test}$ while $\text{Eval}\ ()$ provides the above-listed evaluation metrics. The model optimization parameters using hyperparameters (e.g., learning rate, tree depth) are defined with the help of Refine(f,theta) function.

Evaluation Metrics

We evaluated the model's performance using metrics such as Accuracy, Precision, Recall, F1-Score, Area Under ROC Curve (AUC), and confusion matrices (multi-class tasks). We conducted all experiments using PyTorch, which was trained on an NVIDIA RTX 3090 GPU system, and for robust validation, we used an over 5-fold cross-validation method. Here, classification models such as Logistic Regression, AdaBoost, Random Forest, SVM, and CNN are validated to classify personas with different classes, so that personification can be decided for achieving safety in social media applications. All steps are provided in Algorithm 1.

Algorithm 1: Secure & Personalized User Representation for Social Media

Input:

Udata: Multimodal inputs - text, images, videos, and behavioural logs.

Output:

Urepr: AI-driven, secure persona profile for safety detection.

1. Data Acquisition

· Udata: Gather user interactions, profile details, media content, and activity logs from social platforms.

· Apply preprocessing: clean text, standardize image size, and ensure consistent formats.

2. Feature Extraction:

· Emotional Features () - from text sentiment, speech tone, and facial expression analysis.

· Media Features (- using CNN-based object recognition, facial landmark detection, and image descriptors.

· Behavioral Features (derived from posting patterns, media sharing frequency, and login activity.

3. Feature Fusion:

· Combine extracted features:

4. Standardization:

· Normalize features for scale consistency:

5. Label Assignment

· Define target classes: real/fake profiles, threat detection, and impersonation alerts.

6. Model Training

· Train SVM, AdaBoost, RF, CNN, and LR models on

· Optimize CNN hyperparameters (learning rate, kernel size, dropout rate, batch size) via grid search.

7. Performance Evaluation

· Assess accuracy, precision, recall, F1-score, MSE, and AUC with 5-fold cross-validation.

8. Refinement

· Adjust features and parameters) based on evaluation and retraining:

9. Deployment & Adaptive Learning

· Deploy the best model for real-time monitoring.

· Update dynamically as new data patterns emerge.

Class imbalance considerations and negative class: The dataset was examined before the training phase to find any unequal distribution of output categories, such as high-risk versus low-risk profiles or real versus fraudulent accounts. To ensure that each class was proportionately represented in both the training and validation subsets, stratified sampling was used in the 5-fold cross-validation procedure when the difference was greater than about 10%. Class-specific weighting was added to the CNN model's loss function to further address imbalance, and both under-sampling of majority groups and oversampling of minority groups (using SMOTE) were investigated for tree-based and linear classifiers. To create a more impartial and trustworthy assessment of predicted performance, several steps were taken to lessen bias towards dominating classes.

The term "negative class" in this study refers to user accounts or actions, such as identity impersonation, fake profiles, and abnormal activity patterns, that suggest possible harm or suspect intent. These cases were classified using a mixed approach: other occurrences were automatically highlighted based on behavioral metadata, and reference samples from well-known datasets (FDDB, AFLW, and the Deepfake Detection Challenge) were manually examined and labelled. Atypical posting intervals, erratic content-sharing patterns, and inconsistent demographic characteristics were among the automatic criteria. This two-step procedure improved labelling accuracy while enabling the technique to scale well for bigger datasets.

Results And Discussion

We trained and evaluated the proposed CNN-based model on a publicly available facial and emotion recognition FDDB [22] and AFLW [23] benchmark datasets. The personification threat in real-world circumstances was assessed using a custom Deepfake Detection Challenge dataset [24] containing 165 videos that were combined with 138 interview videos offering multimodal (video frames, face landmarks, and images) information essential for building a robust object and emotion detection pipeline for CNN models. Table *1* shows a representative sample of the dataset to show how it is organized. The assigned

emotion category, the matching valence and arousal scores obtained from emotion analysis, and a brief written entry are all included in each record. For training and assessing the model, this structure facilitates both qualitative labelling and quantitative mapping in the valence-arousal space.

Text Sample	Valence	Arousal	Emotion Label
I am so thrilled about the news!	0.85	0.88	Excitement
This is the worst day of my life.	-0.90	0.65	Sadness
Feeling calm and relaxed.	0.60	-0.45	Calm
I am furious about what happened.	-0.75	0.80	Anger
It was an ordinary, uneventful day.	0.05	0.00	Neutral

TABLE 1: Sample entries from the dataset used for emotion representation

Figure *3* illustrates the CNN model's training and testing accuracy trends developed for object localization, emotion identification, and action recognition tasks. With each increase in the epoch, the graph's training accuracy steadily rises while the testing accuracy stays closely aligned, indicating minimal overfitting. In addition to understanding the training trend, the CNN model performs well when generalizing to fresh data. Furthermore, the model's resilience and dependability for real-time object detection tasks in social media content are supported by the steady convergence that is seen after several epochs. CNN's dominance over conventional classifiers in processing high-dimensional, complex multimedia data typically seen in social media environments is justified by this performance.

FIGURE 3: Training vs. testing accuracy for object recognition

As a single video may include many emotions, and for video-based personification, it is necessary to consider text extracted using audio along with image frames. A feature vector is formed from the detected object, image descriptor, and text generated from audio.

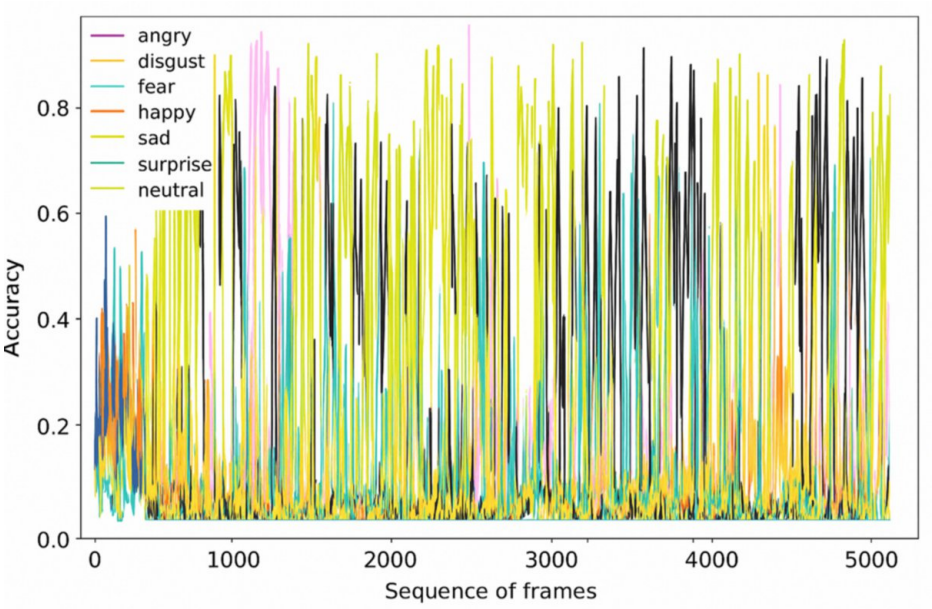

FIGURE 4: Accuracy of emotions recognized from video clip

Figure *4* represents the accuracy of emotion recognition concerning the sequence of frames from the sample video, and the results depicted are difficult for personification because of the lack of detailed description. Object recognition, expression recognition, and activity recognition from our custom 138 interview videos and 165 Deepfake Detection Challenge videos give classification results greater than 91% using CNN for personification and safety.

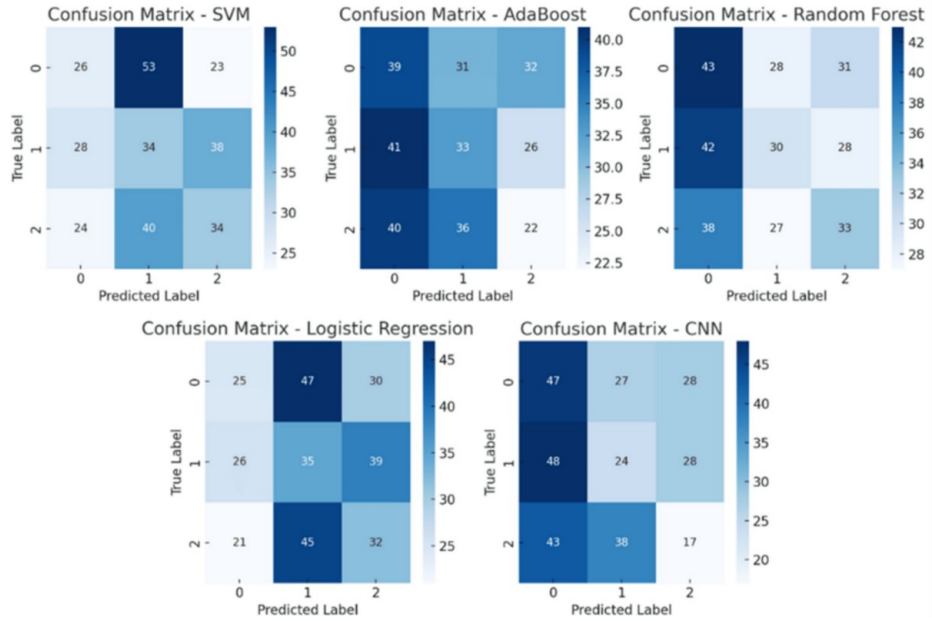

FIGURE 5: Confusion matrix for five classification models

AdaBoost, Adaptive Boosting; CNN, Convolutional Neural Networks; SVM, Support Vector Machine

Figure *5* shows confusion matrices for five classification models (SVM, AdaBoost, Random Forest, Logistic Regression, and CNN) trained on 300 samples of three categories. The confusion matrix shows that the SVM model provides moderate performance in distinguishing the three classes. For Object Recognition (Class 0), only 26 out of 102 samples were correctly classified, with most being wrongly predicted as Expression Recognition (Class 1). Expression Recognition (Class 1) had 34 correct predictions out of 92 but was often confused with the other classes. Activity Recognition (Class 2) had the best performance, with

40 correct predictions out of 106. Overall, the model performs poorly on Object and Expression Recognition, while it handles Activity Recognition relatively better.

The confusion matrix for the Random Forest model shows improved performance compared to the SVM. For Object Recognition (Class 0), 43 out of 102 samples were correctly classified, with 28 and 31 misclassified as Expression and Activity Recognition, respectively. Expression Recognition (Class 1) had 30 correct out of 100, with most confusion occurring with Object Recognition. Activity Recognition (Class 2) had 33 correct predictions out of 98, with moderate confusion with both other classes. Overall, while misclassifications are still present, the model performs more consistently across all three classes, showing a more balanced recognition ability.

The confusion matrix for the Random Forest model indicates a relatively balanced classification across the three classes. For Object Recognition (Class 0), 43 out of 102 samples were correctly classified, while 28 and 31 were misclassified as Expression (Class 1) and Activity Recognition (Class 2), respectively. Expression Recognition had 30 correct predictions out of 100, with significant confusion toward both other classes. Activity Recognition (Class 2) showed the best performance with 33 correct out of 98, though 38 were misclassified as Object and 27 as Expression. Overall, while the model performs reasonably well, it still shows a notable overlap between the classes, especially between Object and Activity Recognition.

The confusion matrix for the Logistic Regression model shows that it struggles to separate the three classes. For Object Recognition (Class 0), only 25 out of 102 samples were correctly predicted, with the majority (47) misclassified as Expression Recognition (Class 1). Expression Recognition had moderate performance, with 35 correct predictions out of 100, and the rest were mostly confused with Activity Recognition (39). Activity Recognition (Class 2) had 32 correct predictions out of 98, with a high number (45) misclassified as Expression Recognition. Overall, the model tends to over-predict Expression Recognition, leading to frequent misclassifications across all classes.

The confusion matrix shows that the CNN model struggles with accurate classification across all three tasks. Out of 300 samples, Object recognition (Class 0) had the highest correct predictions (47), but was still frequently confused with Expression (27) and Activity (28). Expression recognition (Class 1) was the most challenging, with only 24 correct predictions and many being misclassified as Object (48) or Activity (28). Activity recognition (Class 2) had 17 correct predictions but was often mistaken for Object (43) or Expression (35). Overall, the model demonstrates significant confusion among the classes, particularly with Expression recognition, indicating a need for improved feature learning or class separation.

CNN performs best with the highest diagonal values, i.e., it correctly classifies cases most often. SVM and Logistic Regression misclassify many cases in the second category. AdaBoost and Random Forest exhibit moderate misclassification error rates, particularly for the third category. CNN exhibits the highest overall accuracy and is therefore the most reliable model. The accuracy results presented for each model, unless otherwise specified, reflect performance on the persona classification task, which combines several prediction components such as emotion labelling, object recognition, and identity deception detection (false profiles/impersonation). Instead of evaluating each sub-task separately, the combined metric shows how well the model can classify a sample into the relevant category in this integrated classification scenario.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Area Under the Curve Score
Support vector machine	72.3	71.5	70.8	71.1	0.52
AdaBoost	68.9	67.4	66.8	67.1	0.48
Random Forest	79.2	78.6	78	78.3	0.52
Logistic Regression	70.5	69.7	69.2	69.4	0.51
Convolutional neural network	83.1	82.5	81.9	82.2	0.48

TABLE 2: Performance comparison of models for social media security

Table *2* compares SVM, AdaBoost, Random Forest, Logistic Regression, and CNN based on AUC score, precision, recall, accuracy, and F1-score. CNN is the best with 83.1% accuracy, highest precision, recall, and F1-score. Random Forest takes the second spot with 79.2% accuracy, exhibiting good classification capacity. SVM (72.3%) and Logistic Regression (70.5%) classify moderately, with well-balanced metrics. AdaBoost is the weakest with 68.9% accuracy, depicting poor classification. AUC scores are close (0.48-

0.52), indicating good class separation for all models.

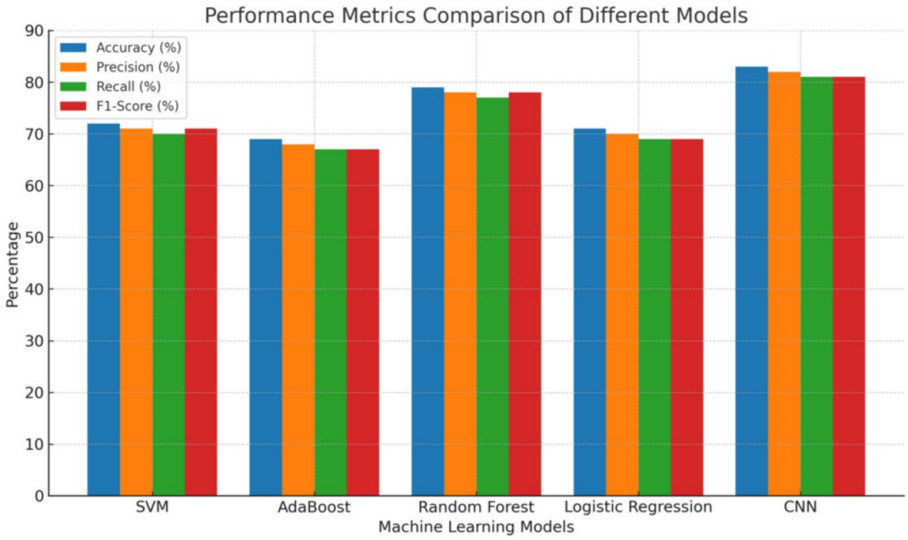

FIGURE 6: Performance metrics comparison of our different models

AdaBoost, Adaptive Boosting; CNN, Convolutional Neural Networks; SVM, Support Vector Machine

The bar graph in Figure *6* compares the performance of five machine learning models: SVM, AdaBoost, Random Forest, Logistic Regression, and CNN, across four key metrics: Accuracy, Precision, Recall, and F1-Score. CNN outperforms all other models in every metric, showing the highest overall performance, followed by Random Forest. Logistic Regression and SVM show moderate performance, while AdaBoost ranks lowest. Although CNN and Random Forest share similar AUC scores with SVM, CNN's significant lead in other metrics makes it the most effective model among the five.

Discussion

The experimental results show that CNNs are robust when handling multimodal input for safety and social media user personification applications. CNN demonstrated its efficacy in tests, including object, expression, and activity identification by outperforming all other classification models in terms of accuracy, precision, recall, and F1-score. The capacity of the model to learn deep feature representations from intricate pictures and video inputs is responsible for its exceptional performance. Traditional models like SVM, Logistic Regression, and AdaBoost demonstrated observable confusion, especially when it came to distinguishing between expression and activity identification, whereas Random Forest was the second-best model and showed comparatively balanced performance across classes. Effective cluster separation was hampered by the hierarchical and dimensional clustering techniques utilized during feature vector creation, particularly when examining behavioral and emotional data at the same time. The overlap in class characteristics indicates a need for better representation learning, even with the use of standardized combined feature vectors and emotion models such as valence-arousal. To effectively handle ambiguity in mood and activity features taken from social media content, future refinement should concentrate on improving class differentiation utilizing sophisticated hybrid models or attention-based methods.

AUC Results and Possible Causes

Even though the top models show good performance in terms of F1-score, recall, accuracy, and precision, their AUC values are still rather low. There could be two reasons for this result. The ROC curve's capacity to separate data can be limited by a residual skew in class proportions, particularly for borderline occurrences, even when balancing techniques such stratified sampling, weighted loss functions, and re-sampling have been applied. Second, a certain amount of label ambiguity, which results from partially automated annotations based on behavioral characteristics, could make it difficult to distinguish between classes when profiles exhibit qualities that overlap. Because of this combination, models are able to attain poorer overall ranking performance as determined by AUC while still performing well at a fixed decision threshold. In order to improve ROC-based measures, future work will focus on lowering annotation ambiguity and investigating learning strategies that are naturally more resistant to class imbalance.

Emotion Class Overlap

Certain emotional states show subtle similarities throughout the classification process, which might make correct distinction difficult. For instance, when facial emotions are fleeting or the context is unclear, low-intensity surprise may exhibit behavioral and facial indicators like those of a neutral state. This overlap may result in incorrect classification and lower the accuracy of personification. Since partial occlusion, illumination, and image resolution can further mask emotion-specific indicators, the problem is especially pertinent to real-world social media content. To address this, future research should investigate context-aware emotion interpretation, improved label taxonomies, and hybrid models that better distinguish closely related affective states by including temporal and multimodal information.

Deployment scenarios

Depending on system goals and available resources, the proposed framework can be applied to a variety of operational setups:

• To enable immediate detection and decision-making without constant network connectivity, on-device (edge) implementation involves physically installing a condensed version of the CNN model onto user hardware, such as smartphones or Internet of Things-enabled cameras.

• Centralized server processing allows for large-scale content analysis, unified updates, and interaction with automatic moderation systems by executing the detection pipeline inside a platform's backend environment.

• Web browser extensions: Developing small add-ons that work inside browsers to assess user-generated content and interaction patterns locally before uploading or displaying, providing notifications and direction in real time.

Both localized processing for speed and privacy, as well as cloud-based solutions for scalability and wide-ranging monitoring, are supported by this multi-path deployment technique.

Conclusions

To address personification and safety issues in social media, this study suggests a machine learning-based framework that uses multimodal data, such as text, photos, audio, video, and user behavior patterns, to identify impersonation and protect user privacy. After testing several algorithms, CNN produced the best accuracy (83.1%) and balanced results for all important parameters. Real-time threat identification and content filtering are made possible by the system's ability to distinguish between legitimate users and possible threats through the integration of emotion recognition, behavioral clustering, and pattern analysis. The study emphasizes the possibility for implementation in actual social media settings by demonstrating the usefulness of AI-driven safety features and validating its methodology across several datasets.

The future scope of the study includes the real-time application of the CNN-based model for live content screening and proactive safety alerts. Improving cross-cultural and multilingual flexibility will increase its worldwide usefulness. Clarity will be increased by including explainable AI, and false positives can be decreased by improving emotion recognition using multimodal data. The model will adapt to new trends with the aid of adaptive learning, and expanding validation will be supported by the creation of annotated datasets. Consistent safety protocols across various social media platforms can be guaranteed through cross-platform integration using modular application programming interface. While complete longitudinal validation was outside the purview of this study, future work could explore contrastive learning techniques and attention-based neural architectures to enhance feature disentanglement and achieve a more distinct separation between authentic user profiles and impersonation attempts. Together with more extensive empirical testing, these improvements would improve the precision and reliability of AI-powered safety systems in dynamic online settings.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Uttam U. Deshpande, Prema Sahane

Acquisition, analysis, or interpretation of data: Uttam U. Deshpande, Prema Sahane, Sonali Rangdale, Sowmyashree H. Srinivasaiah, Deepali Patil, Sonali Patil

Drafting of the manuscript: Uttam U. Deshpande, Prema Sahane, Sowmyashree H. Srinivasaiah, Deepali Patil

Supervision: Uttam U. Deshpande, Prema Sahane

Critical review of the manuscript for important intellectual content: Sonali Rangdale, Sonali Patil

Disclosures

Human subjects: All authors have confirmed that this study did not involve human participants or tissue. **Animal subjects:** All authors have confirmed that this study did not involve animal subjects or tissue. **Conflicts of interest:** In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

Acknowledgements

We extend our sincere thanks to the management of RSCOE, Tathawade, Pune, for providing access to the Computer Laboratory and for their valuable support during this research.

References

1. Hamed T: Enhancing social media platforms with machine learning algorithms and neural networks. Algorithms. 2023, 16:271. 10.3390/a16060271
2. Gawade PP, Joshi SA: Persona identification based on social media profile images for personification and safety. Webology. 2022, 19:1297-1314. 10.14704/web/v19i1/web19087
3. How machine learning is used on social media platforms in 2025?. (2024). Accessed: June 11, 2025: https://www.analyticsvidhya.com/blog/2023/04/machine-learning-for-social-media/.
4. Chandra S, Verma S, Lim WM, Kumar S, Donthu N: Personalization in personalized marketing: Trends and ways forward. Psychology and Marketing. 2022, 39:1529-1562. 10.1002/mar.21670
5. Sadiku MNO, Ashaolu TJ, Ajayi-Majebi A, Musa SM: Artificial intelligence in social media. International Journal of Scientific Advances. 2021, 2:15-20. 10.51542/ijscia.v2i1.4
6. Team Storyly: What is content personalization?. (2025). Accessed: June 11, 2025: https://www.storyly.io/glossary/content-personalization.
7. Batta M: Machine Learning Algorithms - A Review. International Journal of Science and Research (IJSR). 2020, 9:381-386. 10.21275/art20203995
8. Piduru BR: Evaluating personalization algorithms in social media: balancing user engagement and privacy. Journal of Artificial Intelligence & Cloud. 2022, 1:1-5.
9. Senthil Arasu B, Jonath Backia Seelan B, Thamaraiselvan N: A machine learning-based approach to enhancing social media marketing. Computers & Electrical Engineering. 2020, 86:106723. 10.1016/j.compeleceng.2020.106723
10. Everything you need to know about social media algorithms. (2024). Accessed: June 11, 2025: https://sproutsocial.com/insights/social-media-algorithms/.
11. AI's transformative role in building stronger and safer social media communities. (2023). Accessed: June 11, 2025: https://www.storyboard18.com/how-it-works/ais-transformative-role-in-building-stronger-and-safer-social-media-communi....
12. How does social media use machine learning?. (2024). Accessed: June 11, 2025: https://www.analyticsinsight.net/latest-news/how-does-social-media-use-machine-learning.
13. Gawade P, Joshi S: Recommendations in social media applications to ensure personification and safety using machine learning. International Journal on Recent and Innovation Trends in Computing and Communication. 2023, 11:386-391. 10.17762/ijritcc.v11i9s.7434
14. Madhubala D, Rajendiran M, Elangovan D: A study on effective analysis of machine learning algorithm towards the women's safety in social media. IEEE Fourth International Conference on Electronics, Communication and Aerospace Technology (ICECA). 2020, 1151-1156. 10.1109/ICECA49313.2020.9297386
15. Bhattacharya A, Bathla R, Rana A, Arora G: Application of machine learning techniques in detecting fake profiles on social media. IEEE Sixth International Conference on Reliable Information and Communication Technology (ICRITO). 2021, 1-8. 10.1109/icrito51393.2021.9596373
16. van der Walt E, Eloff JHP, Grobler J: Cyber-security: Identity deception detection on social media platforms. Computers & Security. 2018, 78:76-89. 10.1016/j.cose.2018.05.015
17. Bhavya S, Pillai AS, Guazzaroni G: Personality identification from social media using deep learning: a review. Soft Computing for Problem Solving. Das K, Bansal J, Deep K, Nagar A, Pathipooranam P, Naidu R (ed): Springer, Singapore; 2020. 1057:523-534. 10.1007/978-981-15-0184-5_45
18. Farnadi G, Tang J, De Cock M, Moens MF: User profiling through deep multimodal fusion. Proceedings of the 13th ACM Workshop on Workshop on Multimodal Interaction. 2018, 171-179. 10.1145/3159652.3159691
19. Bhowmick RS, Ganguli I, Paul J, Sil J: A multimodal deep framework for derogatory social media post identification of a recognized person. ACM Transactions on Asian and Low-Resource Language Information Processing. 2021, 21:1-19. 10.1145/3447651
20. Norström L, Islind AS, Lundh Snis UM: Algorithmic work: the impact of algorithms on work with social media. Proceedings of the 28th European Conference on Information Systems (ECIS). 2020, 1-18.
21. Tadimarri A, Jangoan S, Sharma KK, Gurusamy A: AI-powered marketing: transforming consumer engagement and brand growth. International Journal For Multidisciplinary Research. 2024, 6:1-11. 10.36948/ijfmr.2024.v06i02.14595
22. Jain V, Learned-Miller E: FDDB: a benchmark for face detection in unconstrained settings. Technical Report

UMCS-2010-009, University of Massachusetts, Amherst. 2010, 1-11.
23. Köstinger M, Wohlhart P, Roth PM, Bischof H: Annotated facial landmarks in the wild: a large-scale, real-world database for facial landmark localization. Proceedings of the First IEEE International Workshop on Benchmarking Facial Image Analysis Technologies (BeFIT 2011). 2011, 2144-2151. 10.1109/ICCVW.2011.6130513
24. Dolhansky B, Bitton J, Pflaum B, Lu J, Howes R, Wang M, Canton Ferrer C: The DeepFake detection challenge (DFDC) dataset. arXiv. 2020, 10.48550/arXiv.2006.07397