Algorithm for Safety Decisions in Social Media Feeds Using Personification Patterns

Prema Pandurang Gawade* and Sarang Achyut Joshi

Department of Computer Engineering, SCTR'S Pune Institute of Computer Technology, Savitribai Phule Pune University (SPPU), Pune, India; Email: sajoshi@pict.edu
*Correspondence: prema.gawade12@gmail.com

***Abstract*—For safety decisions in social media applications, it is necessary to classify personification patterns. The paper proposes using video material to apply machine learning to select, and extract significant feature qualities and grasp the semantics of feature space connection to comprehend the personification of a certain user. The feature traits are based on a computer vision-based approach and a natural language-based approach. A strong belief is calculated from language descriptions and persona traits. These traits are then used to determine the overlap of feature space using various ML algorithms to deduce the intrinsic relationships. The proposed goal is validated by this algorithm and user personification is an important aspect that can be captured through video analytics. Using this personification-based method, better decisions can be made in the given domain space.**

***Keywords*—persona identification, safety, pattern, machine learning**

I. Introduction

Machine learning helps computers to imbibe a natural learning process to solve difficult problems which look very difficult for computers to solve. It includes tasks like natural language, computer vision, and encoding sensory capabilities. This is made possible due to recent advancements achieved in compute scale, storage scale, and availability of exascale of data due to web2.0 standards and data platforms mushroomed in the digital space. These advancements brought high chances of achieving natural intelligence close to human intelligence through several enhancements like Deep Learning, Spherical Confidence Learning Architectures, Boltzmann Machines, and Autoencoders [1–3]. Convolutional neural networks in recent years proved a lot of capability in processing video-based data understanding. The first of its kind of work was demonstrated by LeCun *et al.* [4] for digits recognition using convolutional neural network form which attracts a lot of researchers to prefer machine learning-based approaches over traditional image processing approaches to semantically understand images and videos better. The architectural layout of CNN's layers is depicted in Fig. 1.

As shown a CNN architecture is a layered form where layers like convolutional, activation function, pooling layer, batch normalization, etc. are stacked one after the other. Different parameters and hyper-parameters are used to transform high-dimensional video feed into low-dimensional vector form, making it simpler for architects to align the subject class to a dense vector.

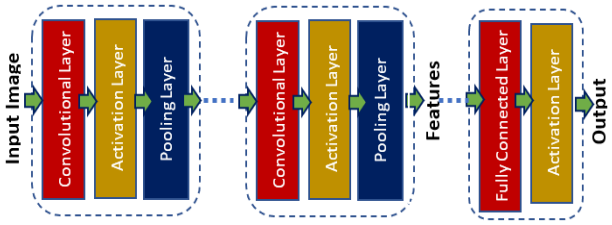

Figure 1. CNN layers.

As the inference is applied with the sigmoid function and then processed using the loss function to understand the deviation of understanding from the ground truth. The objective function then is set to minimize this gap. To achieve this objective CNN uses the backpropagation approach where the derivation of the error function concerning the change in the parameters is calculated which is then passed through all the previous layers till the first layer using the Markov Chain hypothesis. This process ensures that the parameters are changed such that successively it will reduce the cost function to gradually improve the approach's overall accuracy. CNN is used in wide range of use cases such as image-based class estimation [5, 6], face recognition [7, 8], video-based tracking [9], road object detection [10], parameter prediction [11], scene localization and labelling [12], persona recognition [13] image data labelling [14, 15], scene object identification [16], audio classification [17] and natural language processing [18], etc.

Recruitment is critical for any organization's success. Employee sourcing, recruitment, retention, employment, and skill development are key facets of a company's HR management. As it is dependent on the quality of workers working together in synergy to achieve organizational targets and prove themselves of company value. The persona is one of the important traits that organizations are

Manuscript received July 18, 2022; revised September 17, 2022; accepted September 28, 2022; published February 27, 2023.

looking at while recruiting candidates. Understanding, analyzing, and defining the persona of an individual significantly acquisition strategies. Additionally, Human Resource professionals should attract the right talent which suits company expectations. To achieve all the expectations video-based personification stands out as the best solution to avoid all types of biases introduced using the traditional onboarding process.

The proposed approach extracts a candidate persona trait. These persona traits mainly focus on the nonverbal and verbal skills of an individual to understand the persona which is a valuable substance for the talent onboarding team. There are a few hypotheses like the interviewee is ready for an interview in front of the camera, an exact match with the job description, gelling well with the team and company culture, reduction in employee satisfaction after onboarding, etc. will be met by applying the automated persona understanding technique which if implemented using artificial intelligence will ensure the best value, less bias and less time to complete. We are using a five-task approach in creating our candidate personas. These tasks include: 1) Objects Recognition 2) Expression Recognition 3) Action Recognition 4) audio recognition and conversion 5) Text processing. The amalgamation of outputs of the above five steps is engineered to achieve Persona Analysis.

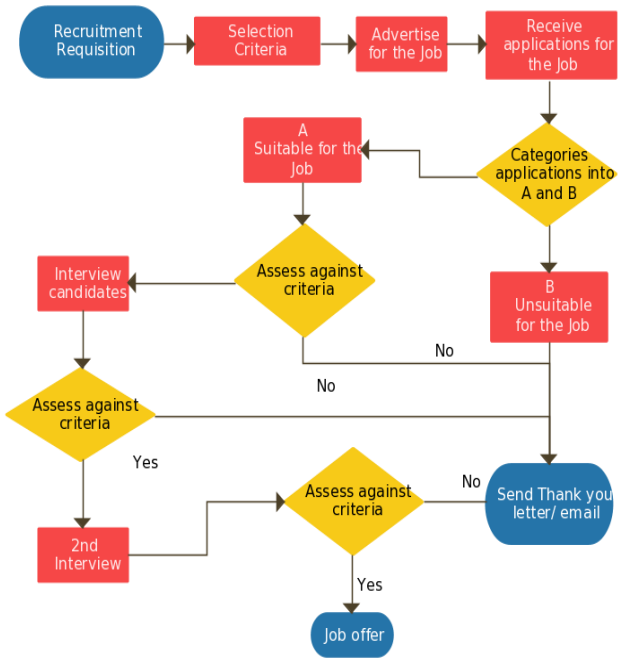

Figure 2. Traditional use case.

Fig. 2 illustrates how understanding a candidate's Persona may significantly enhance a company's recruitment efforts, output, job satisfaction, and overall atmosphere. Of late, a lot of organizations have started leveraging personas for their hiring initiatives. Personas can be used in varied areas of talent acquisition like Sourcing, Recruiting, Employer Branding, Candidate Experience, Training, Digital, and Social Media strategy, Recruitment Marketing, etc.

A. What is Computer Vision?

Processing video data with the aid of computer vision enables machines to comprehend scenes in the visual world. Further, the capabilities are understanding the content, localizing the objects, performing pixel level classification, segmentation, boundary analysis, etc. The overall approach is to use different scale filters to convert the high-dimensional embedding vector into low-dimensional vector space without losing the information. This allows computers to understand the underlying distribution of data and derive the nonlinear relationship between vector semantics and class vectors.

B. What Is Persona Understanding?

A persona is how a person portrays himself or herself to the world, the character features that others see in this person, which may or may not match their true personality. Understanding persona is useful to provide a focused solution in the onboarding solution space. It helps in attaining employability criteria, marketing strategies, and other target-focused solutions. Personas are specific archetypes that are required to understand and identify the right talent from the given pool of audience. A unified vector representation which is a representation of important feature space helps to create a single identity for an individual. A persona has a description representation in vector embedding form where the vector embedding takes a close location in n-dimensional space with other vector embeddings generated for pupils having a similar type of persona. Also, a persona description representation is represented using a natural language-based vector which moreover shows the verbal semantic similarity to understand the similarity of persona.

C. Business Understanding/Problem Statement

Recruitment and skill development are key parameters of a Company's inherent human resource policy. Hiring good people, job satisfaction, reduction in attrition, and maintaining the overall culture of an organization is important. To ensure all of these dimensions, understanding the candidate's persona is important. With the advent of today's digital workforce and agile environment, a lot of organizations started adopting persona-based hiring initiatives in different parts of the human resource department.

Prior studies in candidate hiring show that unbiased, accurate, and adaptive knowledge about candidate persona is vital and important for any organization. Additionally, attributes like facial expression, eye-to-eye contact, and audio-based dialog describe the quality of interpersonal communication [19, 20]. Language traits, Audio traits, and behavioral traits captured from the video feed help understand the verbal and nonverbal behavior of an individual. These all features can be processed as a part of the intelligent video analysis process to strive for intelligent insights which are unbiased most of the time. Quantification of direct and indirect traits captured from video feed to assess the persona in a job interview has not been explored until recent days. In this work, an improved video analytics-based solution help to capture the persona

of an individual in a job interview, and refereeing to HR can take better selection decision.

D. Defining the Candidate's Persona Has Many Benefits

Below are some of the few important reasons to prioritize understanding the persona of a candidate. 1) To reduce the turnover rate, analyzing a candidate's persona is important. 2) Minimize the number of people leaving the organization. 3) Find better suitability with a job description. 4) Creation of analytics insight which will help to hire better talent. 5) Improved candidate experience. 6) Unbiased decisions. 7) Leads learning and development-related improvements. 8) Faster, efficient, and accurate onboarding process. 9) Lower cost. 10) Branding, reputation, and ethics.

II. Literature Survey

There are different ways to process different types of media like matrix, text, images, video, time-stamped events, etc. The more relevant for this work is video data which will be used as a source to process using the artificial intelligence-based technique [21]. Convolutional-based approaches are a great tool to process, understand, and transform video data, which is our main goal in this work. Using this capability, the identity, behavior, and expressions of an individual can be understood and analyzed. NLP decodes the natural form of language used by humans such that the transformed representational form is useful to understand the language. NLP is a branch of AI that additionally takes the domain overlap of linguistic theory and computer science together. Basic NLP tasks include normalization, standardization, parsing, creation of lemma/stems, word sense ambiguity detection, and POS tagging. The Myers-Briggs Personality Type Indicator (MBTI) shown in Fig. 3 helps to analyze the persona of an individual across four dimensions.

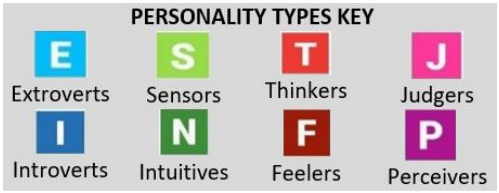

Figure 3. Personality types key.

There are in total 16 different personality types and combinations of those to categorize the profile. The objective of this work is to process an interview video and recognize the person (interviewee) in the video and predict his/her persona based on the answers. The program accepts an interview video as input. The interviewee's face is extracted from the input video and face embeddings are obtained using pre-trained Open Face ML models. The face embeddings are passed to K Nearest Neighbours (KNN) model where the input embedding is compared with the known face embeddings using K-Nearest Neighbours (KNN) algorithm to recognize the interviewee. The interviewee's audio is extracted and Google Speech-to-Text API is used to convert this audio to text. The interviewee's text undergoes NLP processing to derive 10 features which are fed to 4 pre-trained Logistic Regression Classifiers [22] for each personality type group. The combined output of the 4 personality groups is used to map the interviewee to one of the 16 MBTI personality types. This paper will walk you through detailed insights on the use case and implementation thereby keeping into consideration the performance aspects.

III. Proposed Method

The research aims at processing feature space under more than one dimension as shown in Fig. 4 like object detection, expression understanding, action recognition, and natural language-based categorization.

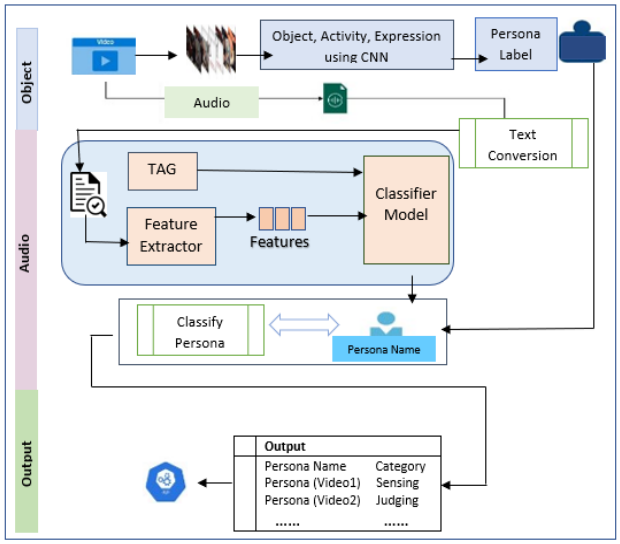

Figure 4. Proposed architecture.

Here, as the dimension space is different, there is no prior approach that works on all the dimensions at the same time. As shown in Fig. 5, the research adopts all these different feature spaces by using variable, adaptive distance metrics while implementing k-means for clustering. This change makes the algorithm more adaptive and less invariant towards slight changes made in the feature space of data. This research opted for a methodology to check the effectiveness of distance metrics toward feature space.

A. Video Frame Analysis Model

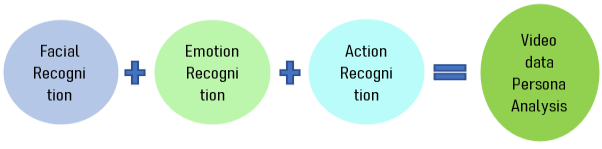

Figure 5. Video data analysis.

B. Object Recognition

Object recognition is a technology-based method of recognizing a human face. As the data distribution is centered around fixed space, there is no need to iterate the weight explorations across the language. This will decrease the training time and ensure optimal generalization. While setting up the parameterization, we

opted for gradient on the very limited area than on the full solution space which effectively captures high density on the localization where there is a high probability of ground truth. This significantly optimizes the performance.

Biometrics are used in an object recognition system to map face traits from a photograph or video. To identify a match, it compares the data to a database of known items. Object Recognition using MTCNN in facenetpytorch and inception Resnet V1 object detection and alignment in unconstrained environments is challenging due to various poses and other types of light effects. Based on recent studies it is evident that deep learning-based approaches outperform traditional approaches irrespective of variances in environmental aspects.

$$L = y_{ij}\ \mathrm{h}(0, \left\|f(x_i) - f(x_j)\right\|_2 - \epsilon^{+}) + (1\text{-}y_{ij})\ \mathrm{h}(0, \epsilon^{-} - \left\|f(x_i) - f(x_j)\right\|_2) \quad (1)$$

y_{ij} = 1 means sample $x_i\ and\ x_j$ are matching, h=max and y_{ij} = 0 is not matching. $\epsilon^{-}, \epsilon^{+}$ control the margins of the matching, not-matching face pairs respectively and f is the feature embedding function.

Multi-Task Cascaded CNN is a new type of CNN that additionally detect landmarks on faces. It is widely used in many different production-level tools for face detection. The architecture is a stacked form of three neural networks which are connected in a cascaded manner. In this approach, MTCNN is implemented in PyTorch with CUDA-ready implementation.

ResNet – A new type of neural network architecture which allows using deeper networks to achieve better generalization performance by throwing out the embedding from early layers to deeper layers by using skip connections.

Inception Resnet V1 – This is one most important works from Google. These convolutional layers are stacked deeper to allow better generalization. There are several versions of this network, each version is tricked by embedding a greater number of convolutional layers, more variety in batch normalization, utilization of three-by-three filters, etc. The advantage is inception block extracts better input embedding at different scales of image dimension. As the depth is increased in deeper layers at different scales, it allows to achieve better accuracy and performance.

C. Emotion Recognition

The emotion recognition approach allows one to understand the sentiment of an individual by examining his/her facial expressions using different image processing and convolutional neural network-based AI approach. The objective is basically to understand the persona of an individual based on image input. As only a single image fails to capture the emotions, it is evident to use the video stream to understand the emotions from a sequence of image frames. The goal is safeguarding the relationships with self and others, Emotion extraction proved to be an important trait towards achieving the same. Emotion and feeling are expressed at six levels namely Anger, Joy, Fear, Sadness, Surprise, and Love.

$$L_i^{landmark} = \left\|\hat{y}_i^{landmark} - y_i^{landmark}\right\| \quad (2)$$

where $\hat{y}_i^{landmark}$ are the landmarks, coordinate resulted from network and $y_i^{landmark}$ is the ground-truth.

Humans express themselves in different ways including speech, gestures/actions, facial expressions, and handwritten text. Emotion Recognition is implemented by using the pre-trained ResNet34 model using Pytorch, FER2013 dataset samples. Transfer Learning with Pytorch is a super useful technique as it saves a lot of computing and training time by adopting the pre-trained model parameters which are trained on similar types of tasks.

There are two major transfer learning methodologies in deep learning:

1. Fine-tuning: This step involves loading and training a pre-trained model. This relieves the network of a load of random initialization.

2. Feature Extraction: Different techniques are there to extract features from text data and image data [23]. The weights of all layers except the last are frozen, and the model is used for training. The output layer is adjusted according to requirements in both ways. The option to add or delete layers based on various parameters is also preserved as a hyperparameter.

FER2013 is a facial database of RGB types images. Each image is of 48 by 48 dimensions. The total number of images is 30000. The labeling is created using a total of seven labels. These seven labels are happiness, neutral, sadness, anger, surprise, disgust, and fear.

D. Action Recognition

The other important aspect to understand the persona of an individual is recognizing and understanding actions performed by an individual. A video stream that contains the interview recording is taken as input to analyze the actions performed. These actions at the outset are one of the important dimensions to understand the persona of an individual. In order to sequence the visual frames, the spatial component is essential. Some actions can be detected in a single image and some are not. In the otherwise cases, local temporal sequence information differentiates actions that are identifiable in one single image and the otherwise. This approach can be improved in the future by considering more complex actions which require a very large number of in-sequence frames to detect and understand it. Though it is a computationally complex approach the same locally applied temporal approach can be useful in correctly identifying the complex actions. As of now, it is not considered in current work due to infrastructure limitations.

$$\widehat{\varphi_i}\ (\omega) = \frac{1}{OP}\left|\sum_{t=1}^{o} u(t) y_i\ (t) e^{-i\omega t}\right|^2 \quad (3)$$

where, O is the number of frames from video recording, $u(t)$ is the temporal window, $y_i(t)$ is the i^{th} data segment.

Action Recognition using Transfer Learning on ResNet34, Kinetics-400 dataset samples. For this, we use a pre-trained dataset containing action videos and then customize it to suit our requirements. Kinetics 400 – is an

action database that provides a list of 65,000 video clips with a total of 400 action labels. Different types of actions are covered like listening, nodding head, jumping, and dancing as well as human-human interactions such as shaking hands and hugging. The organization of the database is made class-wise, where a class has been assigned 400 video clips. The labeling of actions is performed using human annotations over 10 seconds duration. Persona-based on the key outcomes from each section described in Table I.

TABLE I. OUTCOME OF EACH STAGE

Stage	Outcome
Object Recognition	Object ID creation for each candidate appearing in the video, create a folder for each ID and create a persona sheet to be passed to the next stage.
Emotion Recognition	Detect the different emotions displayed by the candidate appearing in the video, tag it against the individual IDs generated in the previous stage, and then pass the output to the next stage.
Action Recognition	Detect the different actions displayed by the candidate appearing in the video, tag it against the individual IDs generated in the previous stage, and then pass the output to the next stage.

In the end, we will have a list of candidates' IDs along with their persona profiles which specify the range of emotions and actions that they used during the interview videos. With this list, we can derive a lot of valuable insights and use them further for the different talent acquisition programs and practices.

Thus, better personality prediction models which imbibe identity recognition, expression understanding, and action recognition are implemented with higher accuracy and reliability. It can be a super useful mechanism to perform persona analysis for our primary use case which is Job screening.

E. Audio Extraction Model

Once the persona identity is predicted, in the next stage Google Cloud Speech API is used to produce text from verbal responses given by candidates. A total of 30 video files, with one hour of recording for each, were transcribed. An n-gram of length 256 is provided as input to API as a speech intent window. In the experiment, there is no audio processing is implemented to separate the interviewees and interviewer's voices. As it is observed that the interviewer's audio portion is significantly less than the interviewee's so few experiments were conducted to extract the question sentences, but after careful analysis, this approach is dropped as there is no impact on the performance and accuracy of the approach. The Google speech API outputs list of words with their respective confidence rating and the vocalization time. The provided confidence ratings which are nothing but the probability score of recognition assigned to each word helped to pick the word with a higher probability score to ensure better accuracy.

F. Natural Language Processing Model

The persona analysis model executes the below steps shown in Fig. 6 to understand the persona of an individual. 1. Features considered for text enrichment (Feature Engineering), remove Reddit subreddit URLs – Spacy Tokenize – Remove stop words from input – Remove spaces, MBTI types – Lemmatise – Tokenization process followed 2. Personality trait Estimation: Across personality traits, classify them in one of the four labels. Where the MBTI dimensions are decided, each dimension is one label. This is trained and tested using a Logistic Regression classifier [24, 25].

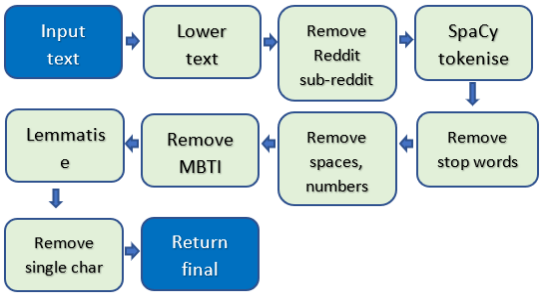

Figure 6. Functional flow.

At the outset, transcription extracted from the video feed is used as input, and then it is designed to predict one of the labels given i.e., Judging, Perceiving Introverted, Extroverted, Intuitive, Sensing, Thinking, and Feeling. The above predictions are combined to map the person to one of the 16 personality types.

G. Algorithm

Algorithm 1: Personification from Video

Input: Video Recordings
Output: Personification
1. Input Video Recording in stored form.
2. Convert into a sequence of frames.
3. Sample one frame from every consecutive second.
4. For each frame
a. Perform Object Recognition
$L = y_{ij}\ h\ (0, \|f(x_i) - f(x_j)\|_2 - \epsilon^+) + (1\text{-}y_{ij})\ h\ (0, \epsilon^- - \|f(x_i) - f(x_j)\|_2)$
y_{ij} = 1 means xi and xj are matching, h=max and y_{ij} = 0 means non-matching.
b. Perform Activity Recognition
$\widehat{\varphi_i}\ (\omega) = \frac{1}{OP}\left|\sum_{t=1}^{o} u(t) y_i\ (t) e^{-i\omega t}\right|^2$
O are total frames and $u(t)$ is the window
c. Perform Expression Recognition
$L_i^{landmark} = \|\hat{y}_i^{landmark} - y_i^{landmark}\|$
5. Extract, and convert the audio recordings into text.
6. Pre-process, and classify the text into 8 identified classes.
7. Y=Train Accuracy
8. Calculate Cumulative Cost Function and regularize.
Return Loss, Y for each frame
9. For each frame, combine the embeddings from STEP:2; 3; 5
10. OUTPUT PERSONIFICATION

H. Flowchart

As shown in Fig. 7 from video recordings vectors of text embeddings and vectors of object recognition, activity recognition, and expression recognition are useful in persona understanding.

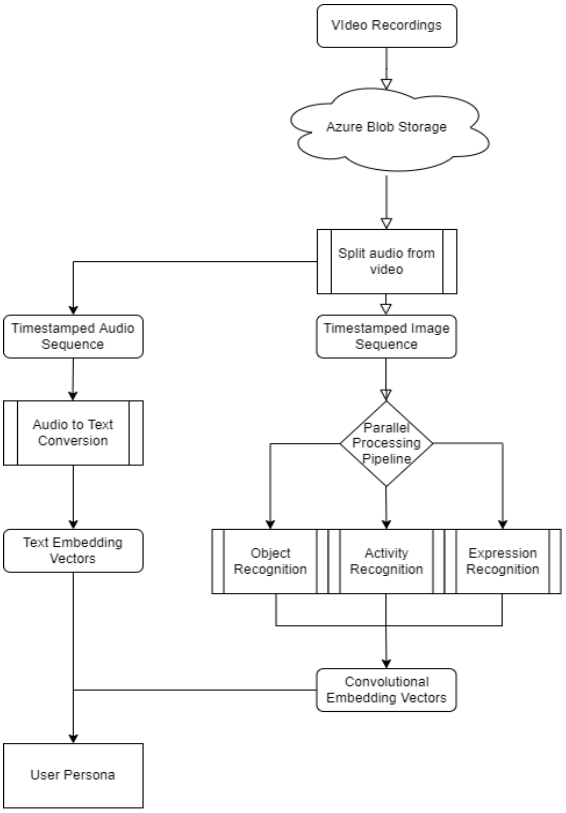

Figure 7. Flowchart to get user persona.

I. Embedding Model

Finding the use case and correct approach from supervised and unsupervised learning is the key to the best model selection. From two models, the first model outputs the embedding vector for video data including personal identification, expression analysis, and action recognition. The second model is used to process the textual extractions provided by google audio-to-text API and then it is processed using natural language processing-based feature engineering. At last, the Logistic Regression and K-means clustering approach is used for Persona Analysis.

IV. Experiments

(1) Input dataset of interview videos as shown in Fig. 8 is provided and the task was to identify the personality traits of the interviewee. Videos start with a conversation and then zoomed in to the interviewee to capture his/her expressions for training another set of images is used for cascading CNN training. Each video has a duration ranging from 40 minutes to 1 hour consisting of more than 10000 frames sampled per video. The videos are clear with almost no noise. Objects can be easily detectable and further processed using the Adaboost approach.

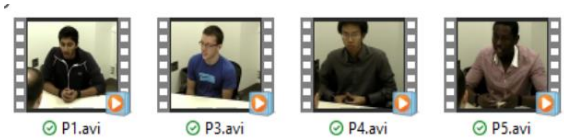

Figure 8. Input video dataset.

(2) Feature Engineering: This step is applied to two different media types, image, and text. For image type of media, the pre-processing is done using initially normalizing, standardizing the pixel data values by dividing it by 256, then it is resized to 256×256×3. As the interview recordings are 1 hr long on average while sampling out the frames from the given video feed the sampling rate is decided as poisonous to ensure better coverage with ground truth universal data available. On text data, different types of normalization methods are applied. Which include steps like removing stop words, converting text into lower case, stemming, lemmatization, and applying vectorization methods like TFIDF, word2vec, and glove to name a few. Once the combination of all these methods is used, it leads to the representation of data for machine learning training purposes.

(3) Training and Validation: Initially we started extracting all the frames from the video and start analysing every frame for facial traits. We end up having thousands of images with most of them having similar outcomes. We then fine-tune our analysis to the frames where there are variations in the traits due to rotation of the head or other movements. We then captured the dominating traits and summarized the same with analysis gathered from the experts. To train the models we used around 70K images apart from the videos. Videos are used to test the outcome of this training. We are providing the character traits of the interviewee to the interviewer to help take a decision.

V. Results and Discussions

The setup is configured on the Google Colab environment which has an Intel Corei7 CPU at 2.4GHz, 2GB GPU, and 25 GB of RAM configuration. Additionally, machine learning training is executed on Facebook Grid Infrastructure. Facebook Grid infra is a vanilla infrastructure cloud that only supports plain infra support with any platform and software support. This helps to get bare hardware for ML training purposes. The software stack used for this work is Python 3.7.1, and MS visual code IDE. With all the relevant packages required for machine and deep learning libraries.

Object localization, then understanding the features from bounding box overlapped region and then analysis of behavioral traits from the input feeds is calculated and then these non-verbal traits are further convoluted with natural language extracts to provide better assumption about persona.

As shown in Fig. 9, all the objects are detected in each frame of the video, and their audio-to-text transcription is achieved with 95 percent accuracy. 98 percent accuracy when we initially tried 10 Epochs and 25K images in both approaches. This is considered overfitting and included more images and also increases the epoch.

It is found that the persona description is provided for each frame in the video. These persona descriptions can be

provided in any form like a set of top category persona, and persona descriptions. As shown, the persona description which is provided by the approach is crosschecked with the mechanical Turk and the individual as well. From the set of this evaluation, it is concluded that most of the time, the persona identified and described by this method achieved higher accuracy and acceptance.

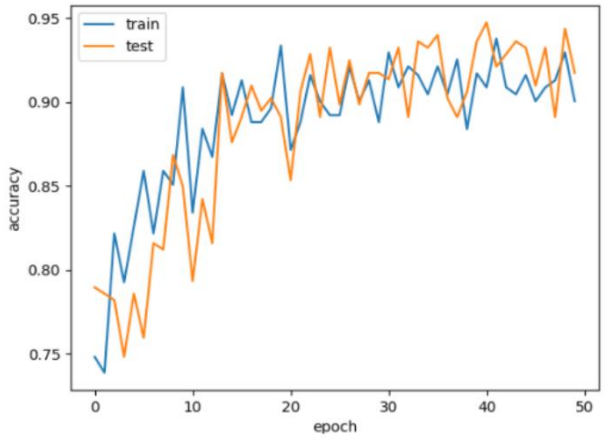

Figure 9. Model accuracy VS epoch.

As shown in Fig. 10, these persona descriptions are further classified into a set of categories shown. This way of categorization ensures that every individual persona can't be categorized as mutually exclusive but it is a collection of more than one category class.

It is evident from these results that providing such percentage distribution across a set of classes makes more impact and use in different use cases based on video-based persona analysis.

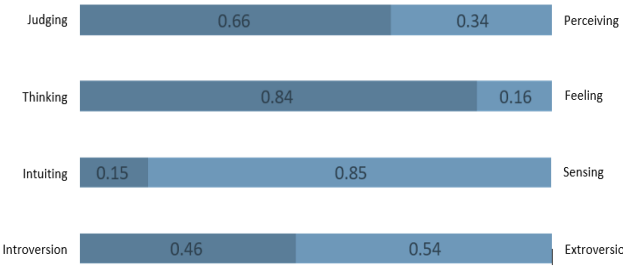

Figure 10. Persona analysis statistics.

Figure 11. Performance analysis of CNN architectures.

A. Performance Analysis

As shown in Fig. 11, the Modified InceptionResnetV4 (MIRCNN) architecture of CNN outperforms with different feature spaces for object recognition, activity recognition, and expression understanding.

Table II shows the performance analysis of the Logistic regression algorithm with 98 percent accuracy used in persona classification and persona analysis, and the K-Mean clustering algorithm with 96 percent accuracy.

TABLE II. PERFORMANCE ANALYSIS

Algorithm	Accuracy
K Means Clustering	96.66 %
Logistic Regression	98.67 %
Convolutional Neural Network	98.45 %

B. Possible Use Case

Recruitment: While recruiting new talent, persona analysis is really important to judge suitability for the job. There are different types of roles like technical, non-technical, functional, etc. For all such categories of roles, the requirements and expectations are different. Even though persona interest may change over time, persona analysis is required for existing employees [26]. Video interview-based personification will greatly help in fixing all loopholes, biases, time lags, and manual interventions. So, this solution is super useful in replacing the traditional onboarding process with intelligent video-based persona analysis.

VI. CONCLUSION

This research results in the identification of a persona based on video data. This design analyses the persona of an individual using objects, and speech traits present in the video feed. Moreover, the text embedding is further processed through the feature engineering pipeline then using logistic regression to categorize the individual into one of the persona classes. Using mechanical Turk, the evaluation is assessed over randomly sampled individuals. This novel research agreed with high confidence. Based on focused persona analysis conducted through automated video analytics, the solution offers improved persona identification to assure lower risk in onboarding new talent.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

The research is conducted under the supervision of S. A. Joshi. Research scholar, Prema P. Gawade has collected a dataset of interview videos and used pre-trained datasets of FER2013, Kinetics-400 for research work. The initial draft of a paper is written by Prema P. Gawade and approved by S. A. Joshi. Both authors approved the final version.

REFERENCES

[1] A. Gopalan, D. C. Juan, C. I. Magalhaes, *et al*., "Neural structured learning: Training neural networks with structured signals," in *Proc. 14th ACM International Conference on Web Search and Data Mining*, 2021, pp. 1150–1153.

[2] D. Alberici, A. Barra, P. Contucci, *et al*., "Annealing and replica-symmetry in deep Boltzmann machines," *Journal of Statistical Physics*, vol. 180, no. 1, pp. 665-677, 2020.

[3] S. Li, J. Xu, X. Xu, *et al*., "Spherical confidence learning for face recognition," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2021, pp. 15629-15637.

[4] Y. LeCun, L. Bottou, Y. Bengio, *et al*., "Gradient-Based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278-2324, 1998.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, pp. 1097-1105, 2012.

[6] R. Madan, D. Agrawal, S. Kowshik, *et al.*, "Traffic sign classification using hybrid HOGSURF features and convolutional neural networks," in *Proc. ICPRAM*, 2019.

[7] Z. Liu, P. Luo, X. Wang, *et al*, "Deep learning face attributes in the wild," in *Proc. IEEE International Conference on Computer Vision, IEEE Computer Society*, 2015, pp. 3730-3738.

[8] I. Adjabi, A. Ouahabi, A Benzaoui, *et al*., "Past, present, and future of face recognition: A review," *Electronics*, vol. 9, no. 8, 1188, 2020.

[9] W. B. Li, M. C. Chang, and S. Lyu, "Who did what at where and when: Simultaneous multi-person tracking and activity recognition," arXiv preprint, arXiv: 1807.01253, 2018.

[10] Q. Fan, W. Zhuo, C. K. Tang, *et al*.," Few-Shot object detection with attention-RPN and multi-relation detector," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2020.

[11] P. Chaudhari, A. Oberman, S. Osher, *et al*., "Deep relaxation: Partial differential equations for optimizing deep neural networks," *Research in the Mathematical Sciences*, vol. 5, no. 3, pp. 1-30, 2018.

[12] H. Zhao, J. Shi, X. Qi, *et al.*, "Pyramid scene parsing network," in *Proc. CVPR*, 2017.

[13] K. Hua, Z. Y. Feng, C. Y. Tao, *et al*., "Learning to detect relevant contexts and knowledge for response selection in retrieval-based dialogue systems," in *Proc. 29th ACM International Conference on Information and Knowledge Management*, 2020.

[14] C. Zimmermann, D. Ceylan, J. Yang, *et al*., "FreiHAND: A dataset for markerless capture of hand pose and shape from single RGB images," in *Proc. IEEE/CVF International Conference on Computer Vision*, 2019, pp. 813-822.

[15] H. Zhang, I. Goodfellow, D. Metaxas, *et al*., "Self-Attention generative adversarial networks," in *Proc. International Conference on Machine Learning*, 2019, pp. 7354-7363.

[16] S. M. A. Eslami and J. Rezende, "Neural scene representation and rendering," *SCIENCE*, vol. 360, pp. 1204-1210, Jun. 2018

[17] Z. Nasrullah and Y. Zhao, "Music artist classification with convolutional recurrent neural networks," in *Proc. International Joint Conference on Neural Network*, 2019, p. 18.

[18] S. Wu, K. Roberts, S. Datta, *et al*., "Deep learning in clinical natural language processing: A methodical review," *Journal of the American Medical Informatics Association*, vol. 27, pp. 457-470, 2020.

[19] S. Song, C. Lan, J. Xing, *et al*., "An end-to-end spatiotemporal attention model for human action recognition from skeleton data," in *Proc. Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[20] J. H. Mun, M. Cho, and B. Han., "Local-Global video-text interactions for temporal grounding," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10810-10819.

[21] M. A. DeVito, J. Birnholtz, J. T. Hancock, *et al*., "How people form folk theories of social media feeds and what it means for how we study self-presentation," in *Proc. CHI Conference on Human Factors in Computing Systems*, 2018.

[22] T. Andreas and S. Jürg., "Actuarial applications of natural language processing using transformers: Case studies for using text features in an actuarial context," 10.48550/arXiv.2206.02014, 2022.

[23] P. Gawade and S. Joshi, "Feature selection for embedded media in the context of personification," in *Proc. Second International Conference on Inventive Research in Computing Applications*, July 2020, pp. 590-594.

[24] A. J. Molstad and A. J. Rothman, "A likelihood-based approach for multivariate categorical response regression in high dimensions," *Journal of the American Statistical Association*, pp. 1-13, 2021.

[25] D. Song, S. Gao, B. He, *et al*., "On the effectiveness of pre-trained language models for legal natural language processing: An empirical study," *IEEE Access*, pp. 75835-75858, 2022.

[26] B. Jansen, S. Jung, S. A. Chowdhury, *et al*., "Persona analytics: Analyzing the stability of online segments and content interests over time using non-negative matrix factorization," *Expert Systems with Applications*, vol. 185, pp. 1-15, 2021.

Prema Pandurang Gawade is a Ph.D. Scholar at PICT, Savitribai Phule Pune University, Pune, Maharashtra, India. She completed her Masters in Computer Engineering from Pune University, India in 2013 and her Bachelors in Computer Engineering in 2004 from Pune University, Pune, India. She has 6 years of industry and 11 years of teaching experience. She taught Undergraduate and Postgraduate Engineering students. She is a member of ISTE since 2013.

Dr. Sarang Achyut Joshi is a Professor at PICT, Savitribai Phule Pune University, Pune, Maharashtra, India. He completed his Ph.D. in Computer Science and Engineering from Bharati Vidyapeeth, Pune, India. He completed Masters in Computer Engineering and Bachelors in Computer Engineering from the University of Pune, India. He works as a Professor in Computer Engineering at PICT, SPPU, Pune, Maharashtra, India for the last 31 years. He was the Chairman of the Board of Studies of Computer Engineering at Savitribai Phule Pune University.