International Journal of Computer Trends and Technology
ISSN: 2231-2803 / https://doi.org/10.14445/22312803/IJCTT-V68I3P115

Volume 68 Issue 3, 74-79, March 2020

Original Article

Automated Training for Job Interviews

Rochelle Cordeiro[1], Anol Kurian[2], Brinel D'souza[3], Brijmohan Daga[4]
[1,2,3,4] *Department of Computer Engineering, Fr. Conceicao Rodrigues College of Engineering, Bandra West, Mumbai-400050, India*

Received Date: 16 February 2020
Revised Date: 31 March 2020
Accepted Date: 03 April 2020

Abstract *- Every individual has to face an interview at least once at some point in their life, but have you ever wondered why one gets rejected even though they satisfy the required educational bars set by companies? Our paper aims at providing a computational framework to analyze supplemental features that go along with subjective knowledge such as facial expressions (e.g., smiles, frowns), language (e.g., word count, textual expression) and prosodic information. The framework aims to be tweakable to suit different companies and their needs by taking technical parameters from previously placed alumni while also checking for coherence with the candidates' CV. It can thus be used to train candidates to better articulate themselves in an interview.*

Keywords *- Convolutional Neural Network, Emotion Recognition, Prosodic Features, Lexical Features, Interview Training*

I. INTRODUCTION

Job Interviews are a stressful part of almost every student's growth which plays a crucial part in securing a good future for oneself. Without any particular exceptions, students or candidates usually scrutinize the subject/topic content relevant to their interview. While we usually presume content is an important parameter in judging a candidate's potential, the reality is far from it. Research conducted by Nalini Ambady and Robert Rosenthal (1993) [8], experimental psychologists at Harvard University, showed that rock-solid impressions are made within the first 20 seconds of an interview. Albert Mehrabian [16], a pioneer researcher of body language in the 1950's, found that the total impact of a message is about 55% nonverbal, 38% vocal and only 7% verbal.

Having established the importance of nonverbal cues in the Human Resource Management industry, our framework aims at providing candidates with a tool to improve their performances in such aspects. The framework looks at tackling this by analyzing facial emotion content, language content (such as pauses, word count, textual sentiments), prosodic content and coherence between candidates' CV, candidates' interview and the Job profile.

The facial component uses a CNN engine to evaluate candidates' facial cues to provide a scrutinized result regarding non-verbal cues. A combination of various text data mining algorithms is combined to put out a detailed report regarding the verbal cues. In the prosodic component, an open-source speech analysis tool is used to inspect the candidates' vocal cues. Finally, with the coherence model, using a similar text data mining algorithm that was applied to analyze verbal cues, an output is generated to show whether the content in the candidates' resume and the interview has any relevance with the companies job description. Since during interviews, HR officials tend to shortlist candidates who show interest in the skills mentioned in the companies job profile, coherence forms an integral part of the system.

Despite significant efforts in this field, automated and objective quantification of our social behaviour remains a challenging problem. Although such technologies have not yet been deployed full strength in the Human Resource Management industry, it is being used as a first screening/filtering process in some prominent companies.

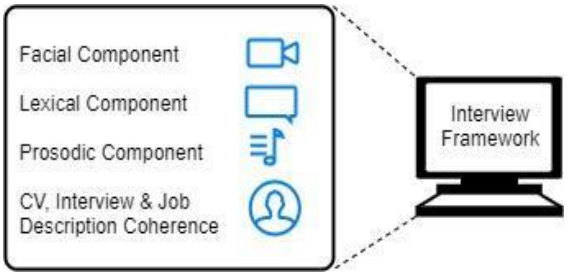


Fig. 1 Interview Analysis Framework. The key components of the framework are facial emotion recognition, lexical analyzers, prosodic analyzer and CV-interview- JobDescription coherence analyzer.

Our framework aims to help candidates clear these filtering processes and also maintain a high level of coherence for real-time interviews, thus highly increasing the possibility of being shortlisted.

II. BACKGROUND RESEARCH

The idea to use CNN for facial features came from a paper by Samer Hijazi, Rishi Kumar, and Chris Rowen (2017) [7] that proved CNN to have the best results in pattern/image recognition. Thus for reduced memory requirements and complexities during training CNN was chosen. From a large set of available datasets, the dataset selected for emotion recognition was found through a research paper by Goodfellow et al. (2013) [11]. The dataset chosen proved to have an accuracy of 63 % with our model.

The research for extracting personality traits was through a paper by M. Fallahnezhad, M. Vali and M. Khalili

[12], through which the basic criteria for analyzing text to extract the personality was known and then on to find the various factors to consider when analyzing text for any individual's personal traits.

The Prosodic features for the framework were guided by Naim, M. I. Tanveer, D. Gildea [1] and M. E. Hoque, together with another paper by M. Kumbhakarn and B. Sathe-Pathak [13] with a detailed knowledge of all the aspects of sound and the various emotions associated with it along with a clear understanding of the relevant tool.

Our research observed many frameworks built to improve the hiring performance for recruiters; hence this framework aims at creating a platform for training the candidates for interviews.

III. PROPOSED SYSTEM

In this section, the proposed system architecture is discussed. It initially describes the major components of the framework along with the flow of the framework and goes on to further describe the components.

A. Overall Description and Flow of Architecture

The system consists of a computer with a working videocam and a working microphone connection. The camera should be of a minimum of 4 megapixels and should be allowed access by the software. The video footage will be relayed to a CNN engine.

B. Detailed Description and Flow of Architecture

a) Facial Component

Emotion recognition is one of the main features for identifying the candidate's traits suitable for the job as well as deciding the results of the interview. For this purpose, the dataset used is the Facial Expression Recognition 2013 (FER-2013) [11] dataset. The data consists of 35,685 examples that are 48*48 pixel grayscale images of faces. The images are categorized into seven categories, namely angry, disgust, fear, happy, sad, surprise and neutral, ranked from 0 to 6, respectively. OpenCV is used to display a square box on the boundaries of the face detected using Haar Cascade Classifiers [14]. According to the face detected in the video, one of the categories with a greater probability is calculated and displayed on the screen.

The process for recognizing emotions is carried on by using the CNN model.CNN model was selected for this system as it has been very useful in saving memory requirements and also reducing the complexities during computation. Considering CNN, any standard convolution layer of a neural network involves input*output*width*height parameters, where width and height are width and height of filter. If we consider an input image of size A*A*M where M is the number of channels and a filter of size B*B*M and the number of operations in 1 convolution operation will be the size of the filter, i.e. B*B*M. Hence the total number of operations will be given by

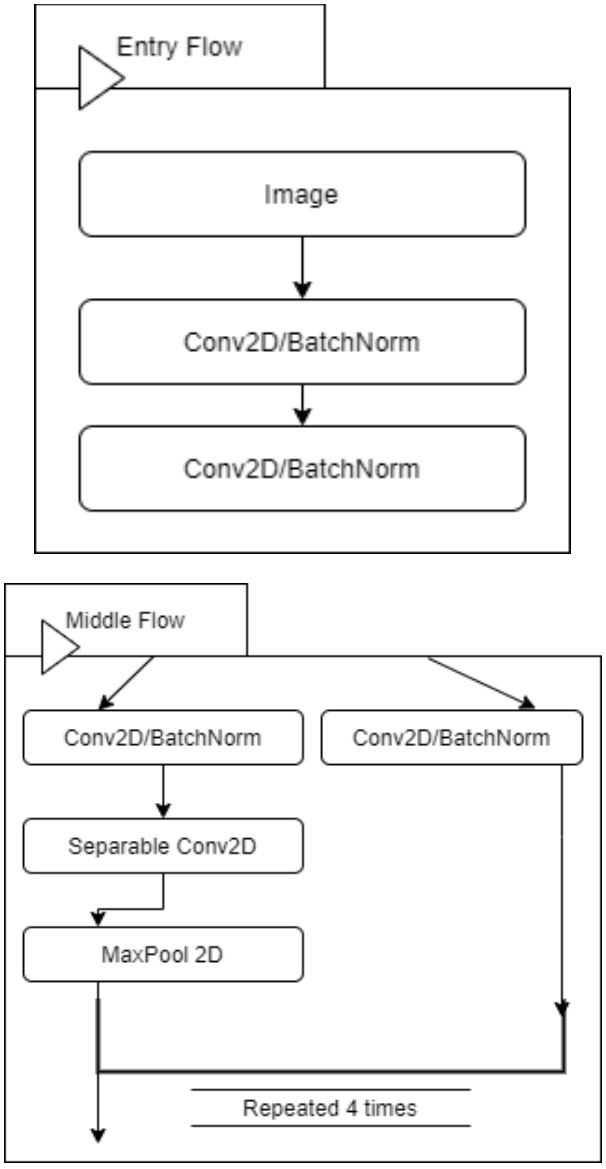
$$Normal\ Convolutional\ Operation = Output * Filter \quad \ldots(1)$$

$$where, Output\ is\ given\ by\ X * X * N$$

For Depth-Wise Separable Convolutions, there are two different layers: Depth-wise and Point-wise. Depth-wise consists of applying convolution to a single-channel, whereas point-wise consists of applying a 1*1 convolutions to the input channels.

$$Ratio\ of\ their\ complexities = 1/N + 1/A \quad \ldots(2)$$

The normal convolutional method resulted in increasing the parameters causing over-fitting. To solve this problem, we have used the Xception architecture.



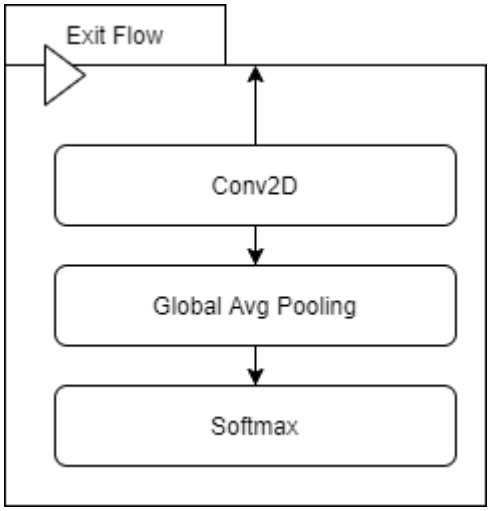


Fig. 2 The Xception Architecture: The data through the entry, middle flow, which is repeated 4 times and exit flow with softmax in the final layer.

This architecture is a Deep Convolutional Neural Network architecture that uses Depthwise Separational Convolutions. The model used in this system is shown in Figure 2.

After the preprocessing and data augmentation of the images have been carried out, the images are trained with a 20% percent split for testing and the rest for the training. The base begins with the image-input passing through the entry flow in which each convolution is followed by batch normalization, and the ReLU activation function that is followed by the middle flow repeated 4 times and ends with the exit flow with Global average pooling computes mean value to reduce the feature map.

Softmax activation is applied at the last layer to convert the output neurons between 0 and 1 to the probability scores.

$$y_i = \frac{e^{z_i}}{\sum_{i=0}^{m} e^{z_i}} \quad \ldots(3)$$

where,

y *is the output neuron*
z *is the input neuron*
m *are the channels*

Also, all Convolutional and SeparationConvolutional layers are followed by batch normalization.

The emotions for each response of the candidate are collected in intervals during the video in the interview process. The values are averaged to obtain the resulting emotion of the candidate of the corresponding question. On a scale of 0 to 1, the emotions are plotted, 0.5 being neutral. Thus the value of the averaged emotion is checked for its equivalent label to get the final result.

b) Lexical Component:

Verbal cues account for the very little impact on a candidate's impression, as little as 7%. Hence we do not concentrate on the content in this component. Our efforts are concentrated on recognizing social intent hidden within the candidate interview. We extract lexical features by using IBM Personality Insights REST API [15], which provides best in class accuracy. The lexical features are analyzed over each individual answer duration, and then a composite picture averaging this analysis is provided for the interview.

This component provides insights on the lines of five big personality characteristics, i.e. agreeableness, conscientiousness, extraversion, emotional range and openness. These characteristics are coupled with two other dimensions, needs and values. The model provides us with 12 characteristic needs and 5 characteristic values. This is used to provide a social intent profile of each candidate.

The API accepts input content as plain text (text/plain), HTML (text/html), or JSON (application/json) by specifying the Content-Type parameter. The default parameter is text/plain.It provides a response in JSON (application/ JSON) or comma-separated values (text/ CSV).

c) *Prosodic Component*

This component reflects the person's tone modulation and thus portrays his social intent regardless of the verbal cues. We extract these features using open source tools My-Voice-analysis built on the PRAAT model [13] for speech analysis. Prosodic features are collected over each single question duration and then compiled together to form a complete emotional profile of the interview.

The tool provides us with various features like standard deviation of fundamental frequencies, minimum and maximum values, pause duration, syllable count and breaks in speech.

d) Coherence Component

The coherence component checks whether the skills required for the job, interview data and resume contents are coherent with each other. This component works with the word count algorithm as its backbone. The company's preferred top 3 technological skills for the job profile they're offering are given to the system as input. A variation of the word count algorithm is used to extract applicable skills from the resume of the candidate where the technical skill relevant to the field in question is stored as a repository in an excel file. This same algorithm is used to parse the interview test data to get the technical skills mentioned during the interview and list them in descending order according to the frequency of using them during the interview. A total of the top 10 such skills with the highest usage frequency are taken into consideration from the resume, and the top 7 skills are taken from the interview data.

Table 1.
Resume/Interview Coherence Calculation Table

Rank	A	B	C
1	50 || 50	30 || 30	20 || 20
2	45 || 42	30 || 30	20 || 20
3	40 || 35	28 || 25	20 || 20
4	35 || 28	25 || 20	18 || 15
5	30 || 21	22 || 15	16 || 11
6	25 || 14	19 || 10	14 || 7
7	20 || 7	16 || 5	12 || 3
8	15 || 0	12 || 0	10 || 0
9	10 || 0	8 || 0	6 || 0
10	7 || 0	5 || 0	3 || 0
11	0 || 0	0 || 0	0 || 0

The cells exclusive of row 1 and column 1 represent (contributionResume)/(contributionInterview) in the format "integer | |integer"

Skill A, Skill B and Skill C are taken as input as the company's preferred skills, with the highest preference being A and lowest being C. The skills mined from resume and interview data will be ranked from highest to lowest frequency, i.e., skills with the highest frequency will have rank 1. Based on what rank Skills A, B and C exist for the candidate, their contributions are calculated from the table above.

The cells represent the contribution of skills to Resume coherence and Interview Data coherence in the format that numbers to the left of the slash show contribution for resume coherence contribution and to the right show interview data coherence contribution.

Calculations⇒

Interview Coherence Evaluation:

α = □□□□□([∑□□□□□□□□□□□□□□□□□□□□□□(□, □, □)] /10)

…(4)

Resume Coherence Evaluation:

β = □□□□□([∑□□□□□□□□□□□□□□□□□□□□□□(□, □, □)] /10)

…(5)

Overall Coherence Evaluation:

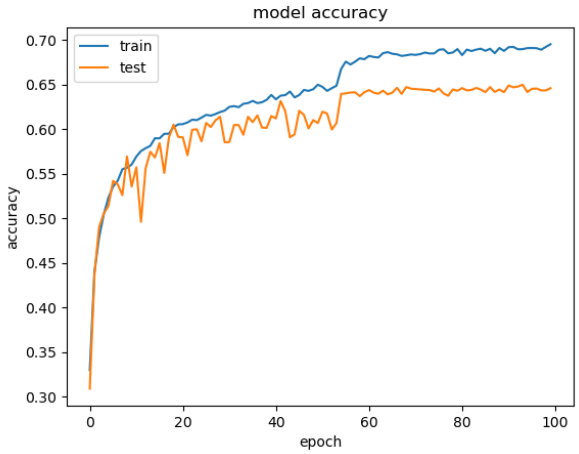
□ = □□□(□, □) ...(6)

An output of an integer digit from 0 to 10 shows the level of coherency between the three topics in question.

IV. RESULTS

The integration of components is still a work in progress, but the models have been deployed and tested on their own independently.

The CNN model for facial cue recognition provides accuracy depending on the quality of surrounding conditions like lighting and background interference. Trained on the FER dataset, which is the go-to data for facial emotion recognition, the model provides a maximum of 63% accuracy, which falls short by 8% from the top-ranking model produced for this dataset on Kaggle. But due to the situation of application intended for this model, an accuracy of 50% is to be considered since it is expected that low camera quality and a high level of background interference will be present.



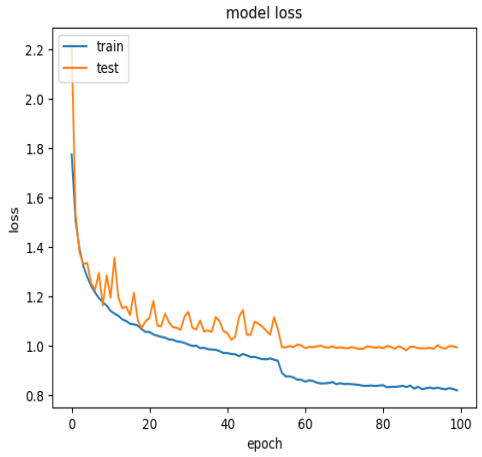


Fig. 3 Model accuracy and loss respectively plotted against epochs up to a count of 100 epochs.

Both lexical and prosodic components use third-party open-source tools that are world leaders in their particular fields. IBM personality insights, which are used for the lexical component, claim to provide a 65% accuracy in their Twitter application but depending on test cases, accuracy also falls as low as 37%. Praat Based tool, which is used for prosodic analysis, claims to provide 70% accuracy in their analysis.

The coherence model provides a ranking mechanism that ranks your coherence in contents between resume and interview. The ranking is an integer between 0 to 10, zero being the lowest and 10 being the highest. Since it's a ranking mechanism, the term accuracy is not relevant to this component. It is as good as the designer deems it to be, a ranking of 10 is provided if the top skills match the job requirements and the ranking is decreased with deviation from the ideal skill rank. The coherency model improves the candidates' performance since the candidate appears more tailored for the job profile.

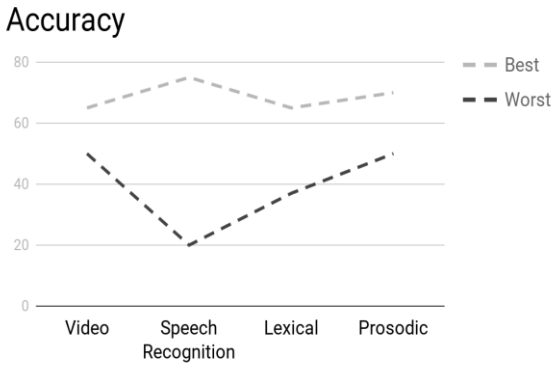


Fig. 5 An estimated approximation of accuracies in best and worst-case scenarios.

Each component provides a high level of accuracy that candidates can depend on to train themselves. The Architecture rivals those that are aimed at recruitment purposes, thus levelling the field for students. By using this architecture model, candidates will be able to better their social intent profiles during an interview, both ai powered and face to face.

V. LIMITATIONS AND ASSUMPTIONS

- Webcam

The webcams accompanying laptops or computers are substandard and thus do not offer high-resolution images, and existing high-resolution web cameras are not feasible monetarily to be used on a student training level.

- Hardware

The RAM specifications and processor speeds may be considered demanding according to Indian standards, namely 8gb RAM and a processor clocking 1500GHz speed minimum is required for the smooth operation of this system.

- Speech Recognition

Python Library for speech recognition is used, so it was assumed no explanation is needed for the same. Only accuracy is taken into account.

- Microphone

High Accuracy Audiophone is required to capture audio that can be legible to the system.

- Lighting

Low lighting conditions coupled with poor camera quality damper the accuracy of the facial component to a significant extent

- Surrounding

The voice recognition model requires low background noise in order to make sense of the audio input.

- Job Description data

The description data regarding Job Profile such as their preferred languages, technologies and platforms, are assumed to be available via alumnus or other external sources

VI. FUTURE WORKS

- Automatic question generation

Use Artificial intelligence to create a chatbot that mimics a human being and create a questionnaire that is based on the candidates' replies. Implement a high level of dexterity in terms of topic flexibility and knowledge.

- High-level facial component accuracy

Current world-class systems have a level of accuracy that roof off at 68%, coupled with the poor camera quality that accompanies most laptops and computer webcams. This reduces the models' dependability. A higher level of accuracy will help the model match human judgment and thus provide candidates with more refined guidance.

- Real-Time Interview

This model can be extended to work behind real-time video interviews and thus provide a second opinion that is void of human presumptions or prejudices.

VII. CONCLUSION

This paper provides a model that is robust and capable of training students for face to face interviews using artificial intelligence, text mining algorithms and prosodic feature extraction to provide overall development and better brace students for real-life situations.

The results demonstrate that while the model rates candidates more critically than HR officials, this critical review helps candidates reanalyze their attitude and sentence formation techniques to better match their profiles.

Apart from facial, textual and prosodic cues, the coherence model provides a solution to the candidates where their resumes, interviews and job profile do not go hand in hand with each other.

The results thus demonstrate that the model provided by this paper can be used in real-life scenarios to train students with social difficulties for placement purposes while still having a scope for further development. This model also leaves scope to be extended as an application in hospitality training, customer management and many socially demanding roles.

VIII. ACKNOWLEDGMENTS

We would like to thank our institute Fr. Conceicao Rodrigues College of Engineering, for their support. We would also like to thank our mentor, Dr B.S. Daga, for his guidance.

REFERENCES

[1] Naim, M. I. Tanveer, D. Gildea and M. E. Hoque, Automated Analysis and Prediction of Job Interview Performance, in IEEE Transactions on Affective Computing, 9(2) (2018) 191-204.

[2] S. Rasipuram, S. B. P. Rao and D. B. Jayagopi, Automatic prediction of fluency in interface-based interviews., IEEE Annual India Conference (INDICON), Bangalore, (2016) 1-6.

[3] L. Chen, R. Zhao, C. W. Leong, B. Lehman, G. Feng and M. E. Hoque, Automated video interview judgment on a large-sized corpus collected online, Seventh International Conference on Affective Computing and Intelligent Interaction (ACII), San Antonio, TX, (2017) 504-509.

[4] H. Suen, K. Hung and C. Lin, TensorFlow-Based Automatic Personality Recognition Used in Asynchronous Video Interviews, in IEEE Access, 7 (2019) 61018-61023.

[5] S. E. Bekhouche, F. Dornaika, A. Ouafi and A. Taleb-Ahmed., Personality Traits and Job Candidate Screening via Analyzing Facial Videos., IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, (2017) 1660-1663.

[6] A. T. Rupasinghe, N. L. Gunawardena, S. Shujan and D. A. S. Atukorale, Scaling personality traits of interviewees in an online job interview by vocal spectrum and facial cue analysis, Sixteenth International Conference on Advances in ICT for Emerging Regions (ICTer), Negombo, (2016) 288-295.

[7] Dachapally, Prudhvi., Facial Emotion Detection Using Convolutional Neural Networks and Representational Autoencoder Units ., (2017).

[8] Ambady, N., & Rosenthal, R., Half a minute: Predicting teacher evaluations from thin slices of nonverbal behaviour and physical attractiveness. Journal of Personality and Social Psychology, 64(3) (1993) 431–441`

[9] R. Chauhan, K. K. Ghanshala and R. C. Joshi, Convolutional Neural Network (CNN) for Image Detection and Recognition., First International Conference on Secure Cyber Computing and Communication (ICSCCC), Jalandhar, India, (2018) 278-282.

[10] N. Jmour, S. Zayen and A. Abdelkrim, Convolutional neural networks for image classification International Conference on Advanced Systems and Electric Technologies (IC_ASET), Hammamet, (2018) 397-402.

[11] Goodfellow, et al. Challenges in Representation Learning: A Report on Three Machine Learning Contests. ArXiv.org, (2013)

[12] M. Fallahnezhad, M. Vali and M. Khalili, Automatic Personality Recognition from reading text speech., Iranian Conference on Electrical Engineering (ICEE), Tehran, (2017) 18-23.

[13] K. Goyal, K. Agarwal and R. Kumar, Face detection and tracking: Using OpenCV., 2017 International conference of Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, (2017) 474-478.

[14] Böhlen, Marc., Watson Gets Personal. Notes on ubiquitous psychometrics., (2016).

[15] Mehrabian, Albert. Nonverbal communication. Transaction Publishers, (1975) .ISBN 0-202-30966-5.