Interview Score Prediction and Analysis Using Audio Features and NLP

Group No: 38

Aditya Jha
Department of CSE
NIT Calicut
Calicut-673601, India
aditya_b180648cs@nitc.ac.in

Ritik Gautam
Department of CSE
NIT Calicut
Calicut-673601, India
ritik_b180630cs@nitc.ac.in

Palash Bajpai
Department of CSE
NIT Calicut
Calicut-673601, India
palash_b180759cs@nitc.ac.in

Abstract — The traditional interview is a test of the personality traits of many hiring managers. However, there are some limitations to the traditional interview method. Besides being time-consuming, it can be difficult for some candidates or interviewers to read candidates because there is no way to know what their voice sounds like under stress or how to properly grade an applicant based on how they responded. The Prosodic feature can help elicit information about emotions, confidence, anxiety, etc, which can help interviewers evaluate candidates. In addition, by extracting recordings from pre-existing audio and applying NLP and other machine learning models, we were able to score the responses of the interviewees. Therefore, in our project, we use machine learning to accelerate candidate evaluation using advanced features and NLP and ultimately rely on reliable data that is less susceptible to human biases thus improving the interviewers' ability to make a hiring decision.

I. INTRODUCTION

In recent years, online video-based interviews have been increasingly used in the hiring process. For example, HireVue, a major vendor in the online video interview hosting market, has reportedly provided its service to many Fortune 500 companies. Conducting online video-based interviews brings many benefits to both interviewers and interviewees, including the convenience of offline reviewing and decision making by human resources (HR) staff, which in turn enables HR staff to assess multiple job applicants in a short time window. Interviewees must adapt their multimodal behaviors, such as speech content, prosody, and nonverbal or facial cues to effectively communicate their qualifications in a limited amount of time. The success or failure of the interviewee's effort is traditionally evaluated by the interviewer, either through personality or quantitative ratings or both according to the need of the company. We can easily interpret the meaning of our verbal and nonverbal behavior during face-to-face interactions. However, we often cannot quantify how the combination of these behaviors affects our interpersonal communications. An emerging alternative to the traditional human-only interview assessment model is to augment human judgment with an automated assessment of interview performance. The style of speaking, prosody, facial expression, and language reflects valuable information about one's personality and mental state. Understanding the relative influence of these individual features and their dependence can provide crucial insight regarding job interviews. These features include facial expressions (e.g., smiles, head gestures, facial tracking points), language (e.g., word counts), and prosodic information (e.g., pitch, intonation, and pauses) of the interviewees. We are proposing a model to design and implement an automated prediction framework for quantifying the ratings of job interviews based on audio features (prosodic, lexical), given the audio recordings. The prediction framework automatically extracts a diverse set of multimodal features (lexical and prosodic) and quantifies the overall interview performance, the likelihood of getting hired, and 14 other social traits relevant to the job interview process. We would be extending the idea of interview score prediction to more like a virtual coach or a tool that would help the candidates to improve them by going through the feedback and suggestions given by the tool. It would be a kind of interactive tool that would synthesize the actual features that our computational framework would need and try giving them suggestions, summary feedback, and detailed feedback on the specific areas that the candidate wants he can go through the feedback and try to improve based on suggestions. We are planning to use prosodic and lexical features to build our framework focusing more on the questionnaire section and using NLP and deep learning to increase the accuracy for the questionnaire section and also it would help the candidates to get a better understanding of the interviews.

II. PROBLEM DEFINITION

To build a suggestion-based interview analysis tool that can help companies to score candidates based on lexical and prosodic features also gives feedback and suggestions to interviewees.

III. BACKGROUND AND DOMAIN DETAILS

In this section, we discuss existing relevant work on nonverbal behavior prediction using automatically extracted features. We particularly focus on the social cues that are relevant to job interviews and face-to-face interactions. As the earlier study showed that motion cues and gesture dynamics plays important role in determining the emotions later this study was further

explored and found to be convincing and more research was done on this area. Later many research was conducted and it was found to be very convincing for determining one's emotions. These emotions were then proved to use for the prediction of the candidate's performance during the job interview. Later many frameworks were developed incorporating many features to predict job interview performance such as audio emotions, facial emotions, etc. This domain particularly involves signal processing, facial coordinates, Action Units used for mapping facial coordinates to emotions. Openface is one such prebuilt tool available. Similarly for audio processing, many prebuilt tools and frameworks are available one of them is the shore framework. Audio features include many features such as spectral features, chroma features, pitch, etc. In addition to that Natural Language Processing can be used to handle the lexical features in contributions to deep learning and machine learning.

IV. LITERATURE SURVEY

- [1] Addresses the challenge of automated understanding of multimodal human interactions, including facial expression, prosody, and language. They used the open-source speech analysis tool PRAAT for prosody analysis, LIWC for lexical features. SVR and LASSO were used as regression models. Ratings predicted by the model are based on social and behavioral skills only.
- [2] Proposed a machine learning-based method to check a candidate's aptitude and personality score based on uploaded CV. The TF-IDF algorithm is used to perform the analysis as a graph in terms of the programming skills on the x-axis and the respective scores on the y-axis.
- [3] It proposes Social Signal Processing (SSP) which provides a general framework of using multimodal sensing and machine perception to analyze human communication including job interviews. BARS rating method was also used. FE method that is highlighted by applying the Neural Network (NN) based doc2vec paradigm to obtain effective visual features.
- [4] SER has also been used by the Convolutional Neural Network (CNN) and CNN Alex Net models. (FACS)Facial Action Coding System use for facial recognition which assigns a numerical value to each facial moment. DNN Model was used for resume parsing and verification.
- [5] Addresses the challenge of HPC- High-performing clusters for prediction of job logs use regression and neural networks models to manage the prediction of new jobs in the job logs so that performance of HPC can be increased.
- [6] This is an automated coach tool developed to provide the interviewee with some insights about the interview. This focuses on three features, expressiveness, response pattern, and acknowledgments. The virtual coach was designed by keeping in mind facial, non-behavioral synthesis Animation of the coach, etc.
- [7] A software tool for real-time fully automated coding of facial expression. It provides estimates of facial action unit intensities for 19 AUs from the Facial Action Unit Coding System (FACS) a, as well as probability estimates for the 6 prototypical emotions (happiness, sadness, surprise, anger, disgust, and fear).
- [8] This also uses nonverbal cues to make a computational framework for hire ability. The basic approach is extracting the video cues and the audio cues to build a baseline model using ridge regression and then using questionnaire data to measure correlation with hire ability score. The R2 value validated the model up to 36 percent.
- [9] The system consists of a speech recognizer trained on non-native English speech data, a feature computation module, using speech recognizer output to compute a set of mostly fluency based features, and multiple regression scoring models which predict a speaking proficiency score for every test item response, using a subset of the features generated by the previous component.
- [10] This paper quantifies the non-linguistic speaking style of engineering school students in practice job inter- views, using features extracted from their vocal tone and prosody. It finds that successful candidates have a characteristic speaking style and these vocal features can be used to build a predictive model of the interview outcomes, with over 85 percent accuracy.
- [11] We used this research document to obtain an audio data set. This article shows how it is impossible to use a single function to determine a person's emotions, and thus requires the use of multiple modal sets of voice and facial expressions.
- [12] This paper worked on how language can reflect the personality style of a person, we used this for getting our dataset for a lexical model.

V. DATASET

Audio dataset: For the audio dataset, we have used Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). For the preparation of this dataset, researchers asked 24 professional actors (12 female, 12 male) to pronounce two identical statements in a North American accent. This produces 7356 files which end up being 24.8 GB of data. Each file is scored 10 times on emotional validity, intensity, and genuineness. Ratings were provided by 247 individuals who were characteristic of untrained adult research participants from North America. A further set of 72 participants provided test-retest data. Since this dataset includes both speech and songs, we had a speech that contains calm, happy, sad, angry, fearful, surprised, and disgusted expressions. Song includes emotions such as calm, happy, angry, sad, and fearful emotions. Each expression is produced at two levels of emotional intensity (normal, strong), with an additional neutral expression. All conditions are available in three modality formats: Audio-only (16bit, 48kHz .wav), Audio-Video (720p H.264, AAC 48kHz, .mp4), and Video-only (no sound). Link: https://zenodo.org/record/1188976#.YhTSFehBy3D

Text dataset: To train our lexical model we used a Stream-of-consciousness data set created during the study of language styles by Pennebaker and King. Their research investigates the authenticity, character structure, and legitimacy of the written language using a computer-based word analysis system. This database used daily texts for 34 psychology students. In this case, 34 students include 29 females and 5 males, with ages ranging from 18 to 67 with an average of 26.4. All these writings form a dataset comprising 2468 records. For each activity, students were expected to write at least 20 minutes a day on a random topic they were given. Data were collected during a two-week summer study between 1993 and 1996. Students were expected to write at least 20 minutes a day on a specific topic for each activity. The scope of personality traits in the sheet is expressed in terms of "y" and "n" representing yes and no to indicate high and low scores. To earn points they used the Big Five innovation techniques to select the best features to score human goals. The Big Five Inventory (BFI) is a reporting scale designed to measure the five major personality traits (addition, acknowledgment, conscience, mind, and openness). Each item contains short phrases and is rated using 5 points on a scale ranging from 1 (strongly disagree) to 5 (strongly agree).

VI. DESIGN AND INPUT-OUTPUT IDENTIFICATION

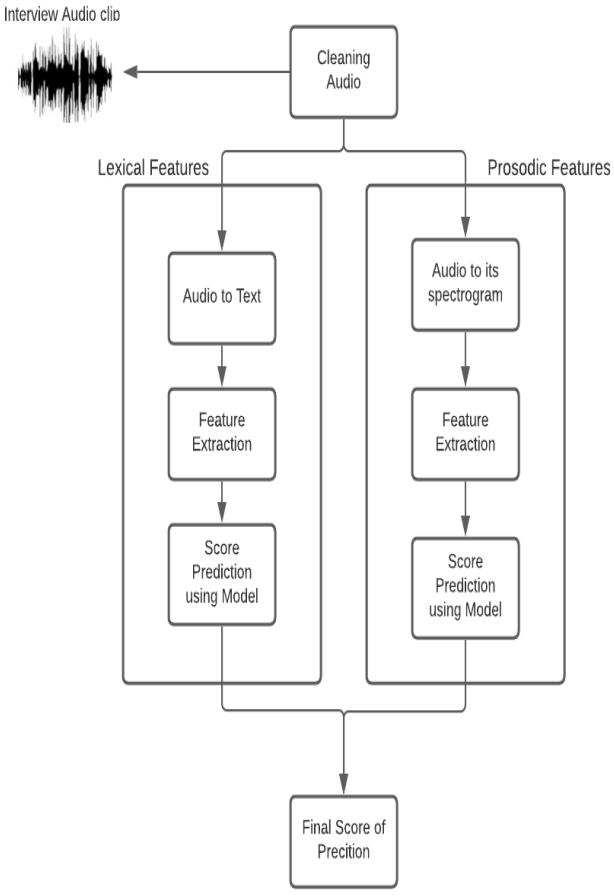

Fig. 1. Implementation approach.

A. Input:

For training purposes, our input to the model will be audio files for the audio analysis model and we will generate transcripts from these audio files which will be used by the lexical analysis model for training purposes. For production websites and apps, the user is presented with a question, and the user answers the question, and the audio from that answer is used to score user interviews in the app.

B. Output:

This is the score assigned to each candidate out of 100. Along with the final score, a behavioral trait score is also given, so you know which factors need further study. A visual analysis of their personality traits is presented. We will also create a feedback system so that applicants know what to do to increase their chances of success in future interviews. So this tool will be helpful for both interview analysis and interview preparation or practical tools.

C. Framework explanation:

We are seeking to construct each component independently and construct a framework that could integrate outputs from each lexical and audio function to generate a rating out of 100. This will assist us to observe a modular technique. So we have constructed two components one for prosodic and one for lexical. Both the components would follow the same general approach converting the data into actual data that we can understand and visualize extracting the features and building a model that would give a score to that candidate out of 100 for now and then merging both the scores based on certain criteria to generate the final score. For the audio part of the framework, we will use a model that provides information about voice quality and voice emotion. Emotions such as anger, fear, sadness, happiness, disgust, and surprise were scored by the audio analysis model. While our lexical model works on the Big Five personality traits. It will mark psychological traits such as openness, agreeableness, conscientiousness, neuroticism, and extraversion (also often spelled extraversion). In this way, our framework uses both emotional and psychological traits for scoring the interview. This framework will also supply guidelines through studying what elements affecting his interview undoubtedly and negatively, as a way to assist applicants to investigate their interviews and boom agree within our version.

D. Final Tool:

Sentence-based interview analysis tools that can help companies evaluate candidates based on lexical and prosody traits also provide feedback and suggestions to interviewees. It also helps you make better hiring decisions by comparing your results with those of other interviewees. Currently, we are only conducting HR interviews, but we plan to expand to all types of interviews.

VII. METHODOLOGY

Audio Analysis model

A. Introduction

Emotions have always been an important part of the personality, and how candidates behave under certain conditions, or questions, are important traits for knowing if they are suitable candidates for the job. Today, voice recognition is used in a variety of areas, from analyzing customer reactions to creating ads that attract viewers, but we are working to use it to select the best candidates for your job. Here for emotional analysis, we will be using the prosodic features of speech. Prosody reflects our way of speaking, especially the rhythm and intonation of speech. Prosodic properties are effective in modeling social intentions. Our system takes into account six emotions: anger, happiness, fear, sadness, surprise, and disgust. We monitor your speech, show which emotions were more dominant in your speech, and provide comparisons with the various candidates our system has been trained on.

B. Theoretical Background

- **Discrete Fourier Transform:** We import the audio signal as wav file which gives us a time-domain representation of the signal. It shows the size (amplitude) of a sound wave that changes over time These amplitudes are not very informative as they only tell you the size of the audio recording. . To better understand an audio signal, it needs to be transformed into the frequency domain. The frequency-domain representation of a signal tells you what other frequencies are present in the signal. The Fourier transform is a mathematical concept that can transform a continuous signal from the time domain to the frequency domain. Audio signals are complex signals made up of multiple "single-frequency sound waves" that propagate together as disturbances (pressure changes) in a medium. So to capture the sound we capture the resulting amplitudes of these various waves. A Fourier transform helps us to decompose a signal into its component frequencies. The Fourier transform gives not only the frequencies present in the signal but also the magnitude of each frequency present in the signal. We will use Discrete Fourier Transform which is a variant of Fourier transform that takes discrete signal as its input and convert it to its frequency constituent.

  For a discrete signal x_n with n=0,……N-1 DFT is defined as

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{\frac{-j2\pi kn}{N}}$$

  Here k represents the frequency domains X_k representing N coefficients.

- **Spectrogram:** A visual representation of the frequency of a specific signal over time is called a spectrogram. In a spectrogram graph, one axis represents time, the second axis represents frequency, and color represents the amount (amplitude) of the observed frequency at a given point in time. We lost information about time after transforming the signal from the time domain to the frequency domain using Fourier transform. Now we don't know at the time what was said if we only use frequency as a function. Therefore, there is a need for a way to find another way to compute the characteristics of the model so that it has a frequency value along with the time generated from the audio signal. This is where the spectrogram comes to the rescue. To calculate the spectrogram of a signal, simply convert the DCT coefficients from powers to decibels and represent them over time. The relationship between power and decibels:

  ydB = 10 * log10(y).

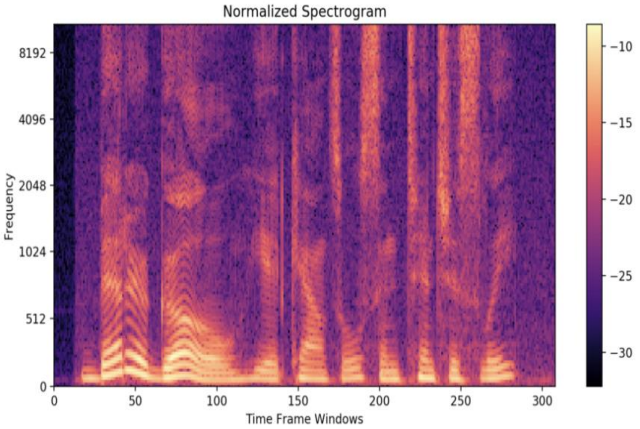

Fig. 2. Spectrogram

- **Log-mel-spectogram:** We used Log-mel-spectrogram for our model. This spectrum fits the Mel scale, which simulates how the human ear works, with research showing that humans don't perceive frequencies on a linear scale. Humans detect differences better at lower frequencies than at higher frequencies, and the mel scale works on this basis. Thus mel spectrograms are used to give our models sound information similar to what humans perceive. The raw audio signal is passed through a filter bank to produce a Mel spectrogram. After this process, each sample is 128 x 128 representing 128 filter banks and 128-time steps per clip. The difference between normal spectrum and logarithmic scale spectrum, both obtained by similar operations, is that while the former shows frequency and decibels over time, the latter shows the relationship between decibels and frequency.

Lexical model

A. Introduction

We learn that each person is different from others in the way we act, the way we speak, the amount we speak, the way we live. Whenever we use words like "talkative", "active", "silent", "creative", etc., we are talking about a person's personality. Thus, personality is the characteristic that distinguishes us from each other. This personality also influences our decisions, our lifestyle, our work, our happiness, our passion for work, and other lifestyle traits. Therefore, selecting candidates who match our work and personality will benefit the company. Therefore, we will evaluate the character of a person. The lexical approach is based on the assumption that the most important personality traits are encoded in natural language words and that structural analysis of these words can lead to the acceptance of personality models. scientifically accepted.

B. Theoretical Background

- **Bag-of-word approach:** The problem with text modeling is that it's confusing and machine learning algorithms like methods prefer well-defined, fixed-length inputs and outputs. Thus our machine learning models cannot run directly on raw text either they prefer to have numerical data. To run a machine learning algorithm, you need to convert a text file to a numeric feature vector. Converts a set of text documents into a token count matrix with several features equal to the size of the dictionary. It was found by analyzing the data (each unique word in the dictionary corresponds to a descriptive function). The simplest and easiest way to count these tokens is to get their frequency of occurrence. The bag-of-words approach helps in extracting features from the text. This method though is quite simple but it can be used to extract different features from the document by knowing the types of words we used and their frequency count. The bag-of-words approach involves two things: A vocabulary of known words and the count of their occurrence. It is called a "bag" since it is just an unordered collection of words irrespective of their position in the document. Thus all information related to the order or structure of the statements is discarded from the analysis. The model is only interested in whether the document has a known word, not the document itself. Thus bag of words help in both preprocessing of the data and also is an important step of feature extraction since from the frequency of the word we will be analyzing the personality of the person. So the use of positive words improves score the personality traits of a person.

- **Big Five Personality traits:** Personality refers to age-old traits and stereotypes that cause people to constantly think, feel, and act in a certain way. As such, personality traits are consistent and enduring characteristic behaviors and emotions. The Five-Factor Model (FFM), which is based on the personality traits of the Big Five, has been developed by several researchers over the past few decades, including Norman (1967), Smith (1967), Goldberg (1981), and McCrae and Costa (1987) The main strength that the FFM model asserts is that it is based on empirical research showing consistency across time, cultural and age groups. It is also considered more structured because the five strokes do not overlap. The traits are:

1. Openness: This trait reflects a person's imagination and insight. People with high values for this trait tend to be creative, adventurous, and very curious. These people are also eager to learn new things and gain new experiences.
2. Conscientiousness: People with strong conscious tendencies are more organized and focused on their goals. Such people like a set schedule and have self-control. They plan, think about how their actions will affect others, and have deadlines in mind.
3. Extraversion: People with this tendency are considered lively, talkative, assertive, and emotionally expressive. They like to hang out with other people, start conversations, and enjoy the fact that they are in the spotlight. Such people can make new friends easily.
4. Agreeableness: This dimension of personality includes attributes such as trust, affection, compassion, kindness, and other prosocial behaviors. People high on this trait show interest in others, love to help others and strive to bring happiness into others' lives.
5. Neuroticism: People with this trait tend to be emotionally unstable. They are sad, irritable, and suffer from anxiety and mood swings. Such people struggle to recover from stressful events.

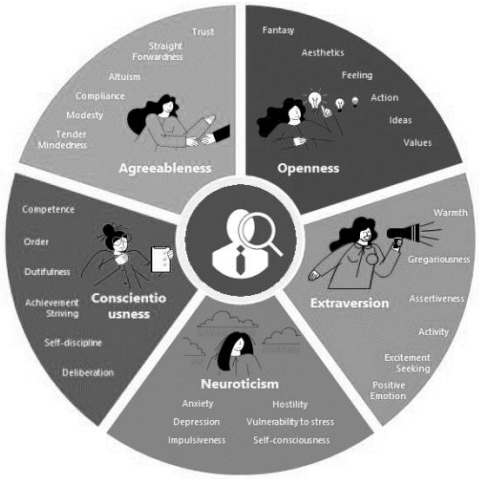

Fig. 3. Big Five Personality Traits

VIII. MODEL

A. *Audio Analysis model*

Input dataset: For the audio dataset, we have used (RAVDESS) dataset. Since this dataset includes both speech and songs, we had a speech that contains calm, happy, sad, angry, fearful, surprised, and disgusted expressions. Song includes emotions such as calm, happy, angry, sad, and fearful emotions. Each expression is produced at two levels of emotional intensity (normal, strong), with an additional neutral expression. This dataset was processed before sending it to the model as input. We converted the audio signals into their respective spectrogram using Discrete Fourier Transform. Now from this frequency vs time graph, we extracted important features using the librosa python library. Librosa is an audio analysis package of Python. Librosa is mainly used when we work with audio data such as making music and automatic speech recognition. Librosa contains almost every utility needed to work on audio data ranging from loading data to feature extraction and manipulation.

RAVDESS								
Emotions	Happy	Sad	Angry	Scared	Dis-gusted	Sur-prised	Neutral	**Total**
Man	*96*	*96*	*96*	*96*	*96*	*96*	*96*	**672**
Woman	*96*	*96*	*96*	*96*	*96*	*96*	*96*	**672**
Total	*192*	*192*	*192*	*192*	*192*	*192*	*192*	**1344**

Table 2: RAVDESS database summary

Working model: Since our input will be a discrete value we chose to work with Time Distributed Convolutional Neural Network. The TimeDistributed layer helps to work with time-series data. It allows using a layer for each input data that means a single model can be applied to several inputs. We implemented Time Distributed Convolution Neural Network using a rolling window which consists of a fixed-size window with time stamps all along our log-mel-spectrogram. Our CNN consists of two blocks consisting of a two-dimensional convolutional layer on which batch normalization is applied, an activation function ("eLu"), a two-dimensional pooling layer, and a dropout. Behind the two boxes are two fully connected dense layers and a "SoftMax" activation function.

Pipeline:

1. Audio signal input
2. Signal discretization
3. Convert to spectrogram
4. Convert spectrogram to mel scale
5. Split spectrogram in windows
6. Predict using a pre-trained model

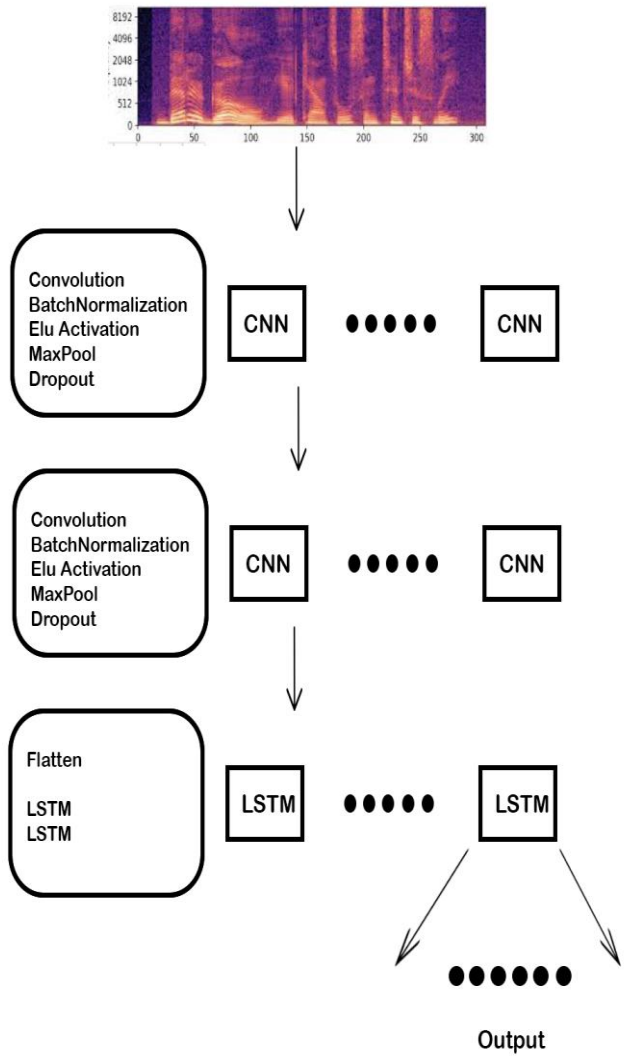

Fig. 4. Audio Analysis Model

Output: A total of 6 emotions are recognized by our audio analysis model namely calm, happy, sad, angry, fearful, surprised, and disgusted. This model will tell which emotion was predominant for your speech and also shows emotion distribution. This will also return a comparison of your emotional distribution with that of others. This emotion distribution will be converted to scores to grade your interview.

Accuracy: 75%

B. Lexical Model

Input dataset: For training purposes, we used a Stream-of-consciousness dataset created during the study of language styles by Pennebaker and King. This database used daily texts for 34 psychology students. In this case, 34 students include 29 females and 5 males, with ages ranging from 18 to 67 with an average of 26.4. The scope of personality traits in the sheet is expressed in terms of "y" and "n" representing yes and no to indicate high and low scores. For input to the model, we will be having text input which we will create from the audio of the answer. Then we will convert this text to a cleaned list of words for which we will do tokenization. After that will use regular expressions to delete unwanted characters from the word list like deleting punctuation characters, converting all to lower case, and removing common words like 'a', 'the', 'is' etc.

Working model: For this, we chose a neural network architecture with a one-dimensional convolutional neural network and a recurrent neural network. RNNs are one of the most promising algorithms used because they are a powerful and robust type of neural network and the only algorithm which consists of internal memory. With internal memory, the RNN can remember important things about the input that it receives which helps it to make more accurate predictions of what will happen next. For this reason, they are the best algorithms for sequential data such as time series. In this way, Recurrent neural networks get a deeper understanding of sequences and their contexts compared to other algorithms. Recurrent neural networks help you to benefit from the sequential nature of information, as opposed to regular neural networks, which assume that the data inputs are independent of each other. We also used the Long Short Term Memory (LSTM) architecture. It is superior to regular CNNs and RNNs by selectively storing patterns for a long period. They do this using memory cells. The final model initially consists of three consecutive blocks consisting of the following four layers: one-dimensional convolution layer, maximum pooling, spatial dropout, and batch normalization. The number of convolution filters is 128, 256, 512 per block, respectively. It also consists of three LSTM cells, each with 180 outputs.

Pipeline:

1. Tokenization
2. Using regular expressions for standardization and deletion of punctuations
3. Converting tokens to lowercase
4. Removal of stopwords (like ‘a’, ‘an’)
5. Padding the sequence of tokens
6. Sending this data to a pre-trained model
7. Predict using a pre-trained model

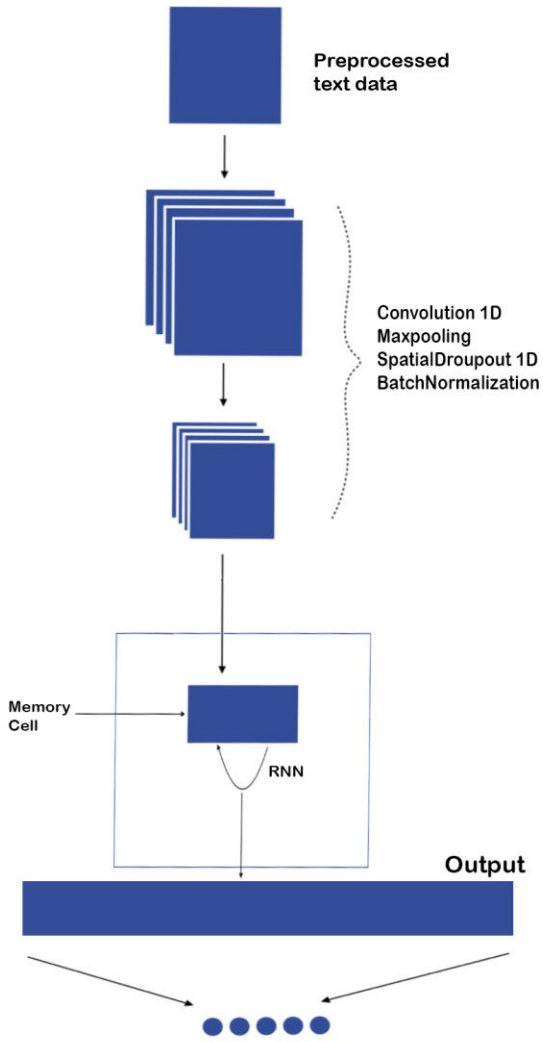

Fig. 5. Lexical Model

Output: Our text analysis model will evaluate a total of five personality traits that can affect your interview namely openness, conscientiousness, extraversion, agreeableness, and neuroticism. The model shows which traits dominate the response and also shows the distribution of personality traits. It also returns the result of comparing your distribution to others. This distribution of personality traits translates into scores for interview scores.

Accuracy: 53%

IX. WORK DONE

- We shifted from MIT dataset to 2 separate datasets for both audio and lexical models.
- We have understood the problem and the requirements to solve the problem efficiently and more accurately.
- We have fragmented our audio files into subparts based on the duration of questions asked, each part contains one question and its answer, and this helps to generate more data, which will help our model to work more accurately.
- We have extracted features from both audio files and stored their values in excel sheets.
- We have made a model for audio analysis using Time Distributed Convolution Neural Network.
- We made a model for lexical analysis using a one-dimensional convolution neural network and recurrent neural network. LSTM is also used to leverage on sequential nature of natural language and to get better accuracy.
- We have modularized our approach.
- Tested each component separately on different test cases.

X. FUTURE WORK PLAN

- Explore more datasets like TESS and SAVEE for the audio analysis model.
- Modify dataset to make it more unbiased to factors like gender distribution.
- Work on the accuracy of all models and test them on different test cases.
- Build a framework to score interviews based on outputs from both lexical and audio analysis models.
- Test each component separately. The components should be loosely coupled and split the features for each component eg : (One component for prosodic features, one for lexical, etc).
- Try to build a suggestion-based tool that takes our model and gives feedback and suggestions to the candidate.
- Make a website and a mobile or pc based application for our model
- Present the final report and give a demo of our work to the panel.

XI. SUMMARY

The scope of this project is largely unexplored and many new ideas can be realized. Interview analysis is used by human resources departments and businesses to determine a person's suitability for the job. This project can even be used by students to train and prepare for interviews. For both audio and lexical features, we use two important human behavior analysis techniques. This project also covers the use of quite different areas of machine learning, so we're going to learn something completely new. I hope this project will be built to company standards and we can use this project in real life.

REFERENCES

[1] I. Naim, D. Gildea, Md. I.Tanveer and Md. E. Hoque, "Automated Analysis and Prediction of Job Interview Performance" - IEEE Transactions on affective computing, VOL. 9, No. 2, April-June 2018

[2] Jayashree Rout, Sudhir Bagade, Pooja Yede, and Nirmiti Patil, "Personality Evaluation and CV Analysis using Machine Learning Algorithm" - IJCSE Vol.-7 E-ISSN: 2347-2693 Issue-5, May 2019

[3] Lei Chen, Gary Feng, Chee Wee Leong, Blair Lehman, Michelle Martin-Raugh, Harrison Kell, Chong Min Lee, and Su-Youn Yoon, "Automated Scoring of Interview Videos using Doc2Vec Multimodal Feature Extraction Paradigm" - ICMI '16: Proceedings of the 18th ACM International Conference on Multimodal InteractionOctober 2016

[4] Supriya Anand, Nihar Gupta, Mayesh Mulay, and Abhimanyu Sherawat "Personality Recognition and Video Interview Analysis" - International Journal of Engineering Research and Technology (IJERT), ISSN: 2278-0181, IJERTV10IS050122 Vol. 10, Issue 05, May-2021

[5] Zhengxiong Hou, Shuxin Zhao, Chao Yin, Yunlan Wang, Jianhua Gu, and Xingshe Zhou "Machine Learning-based Performance Analysis and Prediction of Jobs on a HPC Cluster" - in IEEE Access, INSPEC Accession Number: 19452885, DOI: 10.1109/PDCAT46702.2019.00053, 12 March 2020

[6] ZGwen Littlewort, Jacob Whitehill , Tingfan Wu , Ian Fasel , Mark Frank , Javier Movellan , and Marian Bartlett "The Computer Expression Recognition Toolbox (CERT)" - in IEEE Access, INSPEC Accession Number: 12007742, DOI: 10.1109/FG.2011.5771414, 19 May 2011

[7] Laurent Son Nguyen, Denise Frauendorfer, Marianne Schmid Mast, and Daniel Gatica-Perez, "Hire me: Computational Inference of Hirability in Employment Interviews Based on Nonverbal Behavior" - in IEEE Transactions on Multimedia, Vol. 16, No. 4, June 2014.

[8] Matthieu Courgeon, Jean Claude Martin, Bilge Mutlu, and Mohammed Ehasanul Hoque, "MACH: My automated conversation coach" - DOI: 10.1145/2493432.2493502, issue: Dec 2.

[9] Zechner, Klaus Higgins, Derrick and Xi, Xiaoming and Williamson, David. (2009). Automatic scoring of non-native spontaneous speech in tests of spoken English. Speech Communication. 51. 883-895. 10.1016/j.specom.2009.04.009.

[10] Vikrant Soman, Anmol Madan Electrical Engineering Department MIT Media Laboratory University of Wisconsin-Madison Massachusetts Institute of Technology SOCIAL SIGNALING PREDICTING THE OUTCOME OF JOB INTERVIEWS FROM VOCAL TONE AND PROSODY Appears: IEEE In'tl Conference on Acoustics, Speech and Signal Processing, Dallas TX March 2009

[11] Livingstone SR, Russo FA (2018) The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE 13(5): e0196391. https://doi.org/10.1371/journal.pone.0196391

[12] Pennebaker, J. W., & King, L. A. (1999). Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology, 77*(6), 1296–1312. https://doi.org/10.1037/0022-3514.77.6.1296