Towards Job Screening and Personality Traits Estimation From Video Transcriptions

Towards Job Screening and Personality Traits Estimation From Video Transcriptions

Yazid Bounab (ID), Mourad Oussalah, IEEE Senior Member (ID) Nabil Arhab (ID), and Salah Eddine Bekhouche (ID)

Abstract—In recent years, natural language processing (NLP) has gained new territory beyond its traditional use in many text mining applications. This paper shows the effectiveness of NLP techniques in assessing the human personality from his/her video transcript. Big-five personality traits or OCEAN model includes five personality factors Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness. This model is used to understand human behavior. The task employs multiple regression analysis to produce a score for each of the personality traits. The developed approach has been tested on the APA'2016 competition dataset from Chalearn V2 Challenge Workshop on Explainable Computer Vision Multimedia and Job Candidate Screening Competition @CVPR17. We also performed an estimation for the job interview of the competition. We achieved an average of 89% in personality trait recognition rate and 89.10% in the Job Interview challenge. The results outperform several state-of-art approaches, demonstrating the feasibility of our approach in this kind of analysis. Our system will open up a new direction in multimedia analysis.

Index Terms—Big-Five personality, deep learning, human behavior.

I. INTRODUCTION

In the last two decades there has been a surge of interest in affective computing and personality computing that seek to automatically recognize and synthesize individual personality [10]. This has been performed either through a multi-modal combination of facial images, speaking style, body movement and/or writing style, or a single modality of the above [58]. In this context, the personality calculus primarily focused on estimating the five personality traits, often referred to as the Five Factors Model (or the Big-Five) [50]: Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness.

The Big-Five mode found to prognosticate many life aspects such as work performance, interpersonal relations, emigration, social beliefs, and well-being [6]. We distinguish two research streams in estimating individual personality prediction [39]. The first one advocates a correct recognition of essential personality traits using self or acquaintance reports that often involves interviews and/or successive observations.

The second stream boils down to the process of recognition of the personality traits of an unfamiliar individual. Typically, computational psychology research primarily deals with this second stream as it attempts to estimate the personality traits from videos and multimedia clips highlighting an individual's behavior.

In this context, there were many studies conducted related to human personality estimation. One may mention the INTERSPEECH 2012 Speaker Trait Challenge [51], which provides an audio dataset and extracted features to enable comparison of computational methods for the Big-Five personality traits. With the proliferation of social media platforms, a growing interest has been noticed in predicting apparent personality traits from social media content. Biel et al. [3] used Youtube vloggers video frames to estimate personality impressions using facial expressions and the Big-Five traits. Likewise, Cristani et al. [11] showed that images in which the users "favorite" on Flickr enables the prediction of both apparent and actual (self-assessed) personality traits of Flickr users. Again, Vernon et al. [58] investigated the effects of several physical attributes, for instance, the head size, posture, and chin-length on person's impressions. The approach is based on approachability, youthful-attractiveness, and dominance through a set of face photographs using factor analysis and a linear neural network to predict impressions. Their findings demonstrated a significant correlation between their prediction results and the actual impression dataset as reported in the ChaLearn Looking at People (LAP) 2016 First Impressions challenge [46].

The ChaLearn LAP'2016 competition sheds light on several computational models tailored for Big-Five personality trait identification. A common approach was to use pretrained deep models fine-tuned on the dataset provided for this challenge. However, other models utilizing semantic assumptions (e.g., separating face from the background, separating various gestures) have also been pursued. On the other hand, since the early age of social psychology, the problem of eliciting an individual's personality traits from his/her textual writing has been explored where a strong correlation between linguistic features and Big-Five personality traits was reported [38]. Hence, many advanced deep learning and (complex) computational models have been put forward recently for this purpose [59]. In this respect, many widely used algorithms such as IBM Waston Personality Insights project [37] make use of Linguistic Inquiry and Word Count (LIWC) and MRC (Medical Research Council) Psycholinguistic database dictionary [60]. Although the outcome and accuracy results are sometimes debatable, the importance of this research trend is widely acknowledged in

both computational linguistic and psychological communities.

Motivated by the results of the recent works and trends in psychology of the human personality identification, this work aims to develop new methodological approach for hybridizing multimedia and textual research in Big-Five personality trait estimation, hinting at a new framework for linking modern computer vision analytic and natural language processing-based reasoning. The methodology pursued in this paper relies on existing audio-textual conversion that transforms the problem of Big-Five trait estimation from video or audio analytic perspective into an estimation problem using textual inputs solely. For the sake of simplicity, in contrast to other studies [21], [64], [27], we restrict to the audio part of the video, which results in disregarding the semantic components (gestures, facial appearance, motion, posture) of the video. Therefore, it is interesting to find out how much such transformation can preserve the quality of the Big-Five personality traits estimation. Additionally, in connection with the ChaLearn LAP'2016 competition, we further used the human personality traits as features to predict the Interview Assessment Value (IAV) for ranking job candidates. In this respect, methods based only on facial expressions do not ensure objectivity for job interview estimation. Similarly, the use of face or gestures can reflect the level of interaction skills only. Therefore, the use of every word of the job applicant's spoken statement is likely to be more promising.

In essence, this research employs multiple regression analysis of transcripts textual data obtained from the video of Apa 2016 dataset also known as First Impressions Dataset [1] [47] from Challenge Chalearn V2 Workshop on Explainable Computer Vision Multimedia and Job Candidate Screening Competition CVPR17.

In overall, this paper advocates five vital contributions.

- We demonstrated the feasibility and performance of the audio-textual transformation for estimating Big-Five personality trait.
- We put forward a new preprocessing pipeline for handling textual patterns raised by this transformation.
- A new deep learning architecture with attention mechanism and bidirectional LSTM layers has been suggested for estimating both personality traits and job interview score.
- We compared the performance of several embedding strategies in the proposed deep learning architecture.
- We compared and discussed the overall results with other state-of-the-art results that were reported in the Challenge.

Section 2 of this paper introduces some background and related literature in the field. Section 3 describes the structure of the used dataset. Section 4 provides a skeleton of the study and performance metrics employed. Section 5 emphasizes the preprocessing stage. Sections 6 and 7 detail the methodology employed and the obtained results, respectively. Finally, the conclusion and perspectives of the work are drawn in Section 6.

II. BACKGROUND AND LITERATURE REVIEW

In this section we introduce the ocean model perspectives and the related deep learning solutions.

A. Social psychology and Big-Five personality traits

1) Background: In the social psychology research field, several theories have been put forward to categorize, explain and understand human personality. Trait theory is one of the approaches that has shown its effectiveness in estimating and comprehending human personality [9] where traits are defined as habitual patterns of behavior, thought, and emotion [26], [16]. In particular, traits are assumed to encapsulate aspects of personality that are relatively stable over time and consistent over contextual, temporal patterns, and influence behavior [59]. The Big-Five personality trait model, also referred OCEAN model initiated from the research in [13] and later expanded upon by other researchers, see [43], [53], [15], and [40]. A brief explanation of the taxonomy of the five personality traits is provided below:

Openness (inventive/curious vs. consistent/cautious).
People with a high value of openness tend to have a wide range of interests (imagination, insight, adventurous and creativity with shallowness and imperceptiveness) indicating excitement and curiosity to learn new experiences. On the other side, people with low openness value are usually much more traditional and may struggle with abstract thinking[48], [24].

Conscientiousness (efficient/organized vs. easy-going/careless)
People with a high value of conscientiousness tend to be organized, disciplined, dutiful, aim for achievement, mindful of deadlines/details and prefer planned rather than spontaneous behaviour. Whereas low conscientiousness is associated with flexibility, spontaneity, carelessness, negligence and unreliability [48], [24].

Extraversion (outgoing/energetic vs. solitary/reserved)
People with a high value of extraversion are outgoing and serve to obtain energy in social situations. Low extraversion (or introverted) tends to be associated with individuals who are reserved and have less energy to share in social surroundings where often introverts need a period of solitude and quiet to renew their energy[48], [24].

Agreeableness (friendly/compassionate vs. challenging/detached)
People with a high value of agreeableness are kind, warmhearted, compassionate, cooperative, trustworthy and helpful by nature. Whereas people with low agreeableness value are usually competitive or challenging people, hostile and selfish, often seen as argumentative or untrustworthy [48], [24].

Neuroticism (sensitive/nervous vs. secure/confident)
People with high neuroticism value are likely to be psychologically stressed with unstable emotions such as anger, anxiety, depression, vulnerability, hypersensitive, insecure, and moody. Whereas low neuroticism value is associated with calm, confidence, emotionally stable and resilience, and sometimes lower inspiration [48], [24].

B. Related works in Big-Five personality traits identification

Traditionally, individual's personality can be manifested through his facial expression while speaking, tone of his voice, gestures, and writing styles [42]. Therefore, many modalities, including text segments, audio and video clips, can be used as input to estimate each of the the Big-Five personality traits. ChaLearn First Impressions Challenge organized few competitions in 2016 and 2017, which quickly became a reference in the field of predicting the individual's apparent Big Five personality traits. Reviewing the participants' submissions at ChaLearn (LAP'16 & 17) Challenge revealed that video and audio modalities are by far the most employed ones. In terms of model architecture employed, the top performing teams in case of either audio or video modality showed a strong tendency towards residual network architecture [22] and convolutional neural networks [32], respectively. Other participating teams have advocated the Long Short Term Memory layer (LSTM) as an audio-video feature representation [55]. While others used an extension to Descriptor Aggregation Networks in CNN model [61].

Unfortunately, text modality in predicting human personality and job screening was not much investigated. Although, the text input was used in the multi-model approaches by some participants. For instance, in [12], the authors employed two models for feature representations: the Bag-of-words model (BOW) and the Skip-thought vectors model of the provided transcribed data. BOW model used the 5000 words of the most frequent words in the transcriptions as a vocabulary to build the feature vectors. While, the Skip-thought vector model used an embedding that describes transcripts as 4800-dimensional mean skip-thought vectors [29] of the transcriptions' sentences. To extract skip-thought vectors from the transcriptions, they used a pretrained recurrent encoder-decoder trained on the BookCorpus dataset [65]. Gorbova et al. [17] used the sentiment score associated with the transcribed video. For this purpose, they used SentiWordNet as a tool to get linguistic features from textual data. This generated a negative and a positive weight for each of the 117000 individual words of the dictionary. This was used together with other lexical features such as minimum, maximum, average and sum of positive and negative weights over a sequence of words. In total, they generated a feature vector of size 8 for each word of the video transcription. Estimating personality trait from solely text segments is known to be very challenging due to inherent limitation of natural language processing tools and the subjective nature of natural language, although several computational linguistics have been put forward for this purpose. Traditionally, the correlation between language style and personality trait has been reported by several scholars. For example, extraverts are found to use repetitive statements with fewer pauses, hesitations and lower type/token ratio as well as a more positive emotion words and a less formal language than Introverts ([38]; [14]; [44]; Neurotics tend to use more first person pronouns, more negative emotions, and less positive emotion words [44]. High conscientiousness people tend to avoid negations, and negative emotion words

[44]. Furthermore, several deep learning models have also been investigated for this purpose. Escalera et al. [12] employed diverse Readability indexes on the transcripts using the Natural Language Toolkit (NLTK). They used 8 different measures as features for the Readability representation: ARI, Flesch Reading Ease, Flesch-Kincaid Grade Level, Gunning Fog Index, SMOG Index, Coleman Liau Index, LIX, and RIX. These measures are initially developed for long texts rather than a few sentences in a video transcription. They also included two other statistical features to overall text representation: total word count and unique words within the transcript. Kampman et al. [25] used pretrained Google News word2vec as an embedding representation to encode all words to preserve the contextual information within the transcription. Next, each transcription generates an embedding matrix to be passed to a CNN architecture. The latter is composed of three convolutional layers that extract tri-grams, four-grams and five-grams slid over the transcription with one word each time. After each convolution, a max-pooling layer is performed to get a final transcription encoding for all three layers whose outputs are fed through a fully connected layer with sigmoid activation to the final Big Five personality traits. Table I summarizes the results of state-of-the-art methods for both ChaLearn 2016 and 2017 challenges.

Besides personality traits, Mark Cook [7] introduced many other evaluation types, such as Mental ability, physical characteristics, knowledge, work and social skills.

III. DATASET

The first impression dataset includes 10K video clips of high definition format extracted from youtube online platform of distinct people. The average time duration of each video is about 15 seconds. Each clip has ground truth labels for each of the five traits represented with a score in the unit interval. Videos were labeled with personality trait factors through Amazon Mechanical Turk (AMT) after enforcing appropriate quality monitoring scheme to ensure reliable labelling. Furthermore, AMT workers added the text modality (transcriptions) to the dataset alongside video clips data. Especially, the video clips were transcribed using the professional transcription service Rev. In addition to the big five personality traits labels, the dataset also includes **job-interview** label that contains the probability that the given individual will be invited to the job interview. In the three phases of the challenge, the annotations and the transcriptions data were stored in separate pickle files, where each file is a single dictionary whose keys correspond to the names of the videos and the values are their corresponding transcriptions/annotations. The dataset is split into three main classes: train, validation, and test with a 3:1:1 split ratio, respectively. The train part includes 6000 videos where the validation and test contain 2000 videos for each. The participants' diversity was respected by including people from several social groups of both genders, different age groups, nationalities, and ethnicity. Table II shows the statistical properties of each label in the dataset.

---

[1]Public lexical resource for sentiment analysis

TABLE I
STATE-OF-ART METHODS ON BOTH VERSIONS OF THE FIRST IMPRESSION DATASET.

Method	Modality	Architecture	Regressor/model	Avg. Person.	Interview	key notes
Gürpinar et al. (2016) [21]	Video Audio	VGG-face, VGG-VD19	ELM	**91.30**	-	Text was not used. Only fusion of Scene, face and statistic features from video
Zhang et al. (2016) [64]	Video Audio	VGG-face, ResNet, DNA/DNA+	Linear Regressor	**91.30**	-	Text was not used. Only fusion of Scene and face. Weak regressor
Subramaniam et al. (2016) [55]	Video Audio	3D-CNN	LSTM	91.21	-	Deep features Hand-crafted features 3D-CNN weak in temporal relationship
Ventura et al. (2017) [57]	Video	DAN+	DAN+	91.16	-	Neither Audio or Text was used.
Güçlütürk et al. (2016a)[19]	Video Audio	ResNet	ResNet	91.09	-	Text was not used, Facial expression not included No special treat for audio
ucas (Ponce-López et al. 2016) [47]	Video Audio	lbptop, hog3d, VGG, AlexNet ResNet	Partial least Square Regressor	90.98	-	Text was not used. Hand-crafted audio features weak regressor
Gürpınar et al. (2016b) [20]	Video	CNN	CNN	90.94	-	Neither Audio or Text was used. Only fusion of scene and face
FDMB (Escalante et al. 2018) [12]	Video	LPQ	Support vector Regressor	87.47	87.21	Week features Hand-crafted features
ROCHCI (Escalante et al. 2018) [12]	Video Audio Text	SHORE Pitch and Intensity ASR transcriptions	GBoost Regressor	90.13	90.18	Hand-crafted Text and Audio features
Bekhouche et al. (2017) [2]	Video	LPQ BSIF	Support vector Regressor	91.16	91.57	Neither Audio or Text was used. Facial expression not included Hand-crafted features
Güçlütürk et al. (2017) [18]	Video Audio Text	ResNet	Ridge Regressor	91.18	91.62	Facial expression not included No special treat for Audio
Kaya et al. (2017) [27]	Video Audio	VGG-face, VGG-VD19, ELM	Random Forest	91.73	92.09	Text was not used.
Li et al. (2020) [34]	Video Audio Text	CR-Net	Random Forest	**91.88**	**92.47**	Classification features based on static class division = 10,

TABLE II
SUMMARY STATISTICS OF THE BIG-FIVE PERSONALITY TRAITS AND JOB
INTERVIEW OF APA 2016 DATASET

10.000 Sample	min	max	mean	Std
Openness	0.0	1.0	0.57	0.15
Conscientiousness	0.0	1.0	0.52	0.15
Extraversion	0.0	1.0	0.48	0.15
Agreeableness	0.0	1.0	0.55	0.13
Neuroticism	0.0	1.0	0.52	0.15
Interview	0.0	1.0	0.5	0.15

IV. METHODOLOGY

A. Overall Methodology

We endorse a natural language processing methodology based approach applied to video transcriptions provided in the ChaLearn LAP'16 dataset. This excludes the recourse to audio or video modality in our methodology. Furthermore, capitalizing, on one hand, on the inherent relationship between the interview score and the personality traits as pointed out by some participants of the Fist Impression Challenge [47], and, on the other hand, on the independence between the Big-Five personality traits where in most psychological studies, the traits were often treated as independent variables, we adopt a two-step strategy. In the first step, we develop a deep learning model that estimates the interview score. While, in the second stage, this model is reshaped to estimate each of the Big-Five personality trait. In the sequel, the quality of the textual data requires further enhancement, which calls for appropriate preprocessing stage that will be detailed later on. Figure 1 shows the generic pipeline employed for human personality traits estimation and job screening interview.

B. Preprocessing

Text preprocessing is the first step in any NLP pipeline, which straightforwardly impacts the processing of other subsequent NLP modules as well. This especially holds for the video transcription of ChaLearn LAP'16 dataset where we noticed at least two important types of anomalies, in addition to standard token predisposition irregularities. First, there is a non-negligible discrepancy in terms of the size of the video transcription provided in the dataset where we noticed few empty transcriptions and some containing only few words. Second, there exists an increasing number of inconsistencies
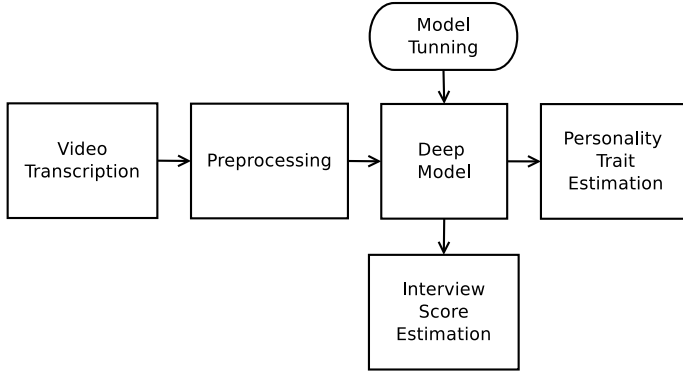
Fig. 1. Generic pipeline

which are rooted back to the miss-positioning of the space character, possibly due to the (professional) automatic video transcription approach employed and the pressure on operators to speed up the manual checking in short time constraints. Therefore, special caution is required to handle these two challenges. We shall refer to the first challenge handling as the high-level pre-processing where individual transcriptions were either maintained or discarded, and the second one as the low-level pre-processing where the content of each transcription is examined for space irregularities.

1) High level preprocessing: Although the handling of empty transcription is straightforward, dealing with small length transcription might be problematic. This is because recognizing personality traits from short text is problematic as well according to psychological research [5], [54]. Therefore, we advocate an approach that discards short text transcriptions and maintains only those which fall within reasonable range. In order to comprehend the threshold beyond which a given transcription is maintained, we conducted a statistical analysis on all provided transcriptions and kept only those cases whose length is more than two time the standard deviation 2σ from the left hand side of the mean value. For this purpose, we first highlighted in Figure 2 the density of transcriptions length over the training set and the selected length range.
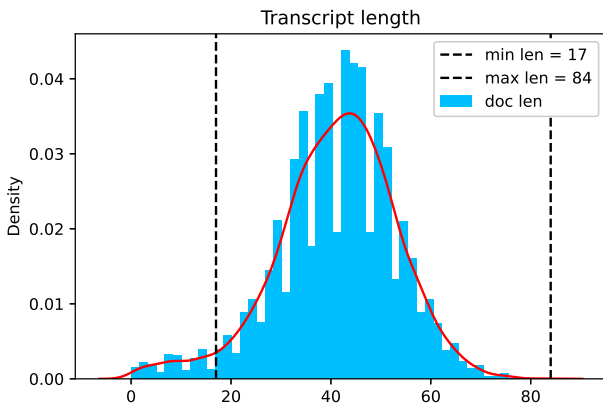

Fig. 2. Density of the length of transcriptions

After this pruning pre-processing activity, the number of

new samples in each set is shown in Table III.

TABLE III
STATISTICS

Set	Old samples	#NANs	#Seq <min length	New samples
Train	6000	11	317	5672
Valid	2000	0	85	1915
Test	2000	7	90	1903

For instance, Table III shows that 18 videos do not have transcriptions. The presence of empty transcriptions can be explained by the quality of the tool used for extracting transcription from a given video. Other reasons include situations where the person facing the camera may be speechless or prefer to use sign language. The table also shows that the dataset includes 492 transcriptions whose size is less than the minimum threshold and are therefore discarded.

2) Low level preprocessing: Prior to handling the space character issue, we first used standard natural language processing approaches for text normalization and cleaning. This consists of the following subtasks.

- Remove chunks that include "[time]".
- Remove URLs and emails.
- Replace combined tokens to separate ones. *"hasn't"* becomes *"has not"*.
- Remove repeated chars in words. *"Maxxx"* becomes *"Max"*.
- Remove Acronyms. *"J.L."*.
- Split words with capital letter in the middle.
- Remove Numbers.
- Remove Punctuation.
- Infer spaces in long words.
- Lowercase texts.
- Remove extra spaces.

The space-character issue comprises two types of irregularities. The first one concerns the presence of space-character in a well-defined word as in "manif estation" where the misplaced space-character splits the correct word "manifestation" into two tokens ("manif" and "estation") that do not match any word in the English dictionary. The second one concerns the absence of space-character resulting in a long chunk as in the token "observationis", which does not match any word in the dictionary, and should be split into two tokens "observation", "is". This second type is found to be dominant in the transcriptions and sounds more complex to handle as well. Therefore, we introduce a new NLP module responsible for inferring spaces in a long chunk of text. This module uses an English dictionary **D** that includes 125k of the most frequent words from a small subset of Wikipedia to identify the words within the string. The words in this dictionary are ordered in the descending order of their frequency in Wikipedia corpus. The restriction to the aforementioned relatively small scale popular dictionary can also be motivated by the fact that spoken language in formal conversation, as for job interviews, is often restricted to a limited dictionary to convey the key messages to the interviewer and score high in communication skills.

[1] http://tinypaste.com/c1666a6b

The problem of space insertion in long chunk can therefore be turned into an optimization problem where a cost function is assigned to individual word (s) resulting from this splitting operation. For this purpose, we first use Zips's law [33], [8]. The latter states that a word ranked **n** in the dictionary has a probability of occurrence approximately **1/(n log N)** where **N** is the total number of words in the dictionary.

To compute the cost function of a word W, we used a combination of its length and its position (frequency rank) in the dictionary D (see Eq.1).

$$Cost(W) = \begin{cases} \log(W_{index}) \times \log(W_{length}) & if \quad W \in D \\ \\ 0 \; if \; W \notin D \end{cases}$$

$$(1)$$

The rationale behind the preceding is to assign a zero cost value to any word W, which is not in the dictionary D. In other words, the best matching words in the sense of maximizing the above cost function correspond to those wordings that achieve a balance between the word length and their rank in the dictionary. This has the tendency to favor popular long wording. For instance, in chunk 'onetwo' would yield "one", "two" as the best configuration. Other configuration ('on', 'etwo') has lower cost value as 'etwo' is assigned zero cost value. In order to check for all configurations, typically dynamic programming can be used. Nevertheless, the process can be simplified by setting a minimum and maximum value of character shift according to smallest and longest words in dictionary D. In overall, for a complexity analysis purpose, it is easily found that for a chunk of length M and let m be the maximum length of words in D, then the maximum cost configuration has a linear complexity $\mathcal{O}(Mn)$. Algorithm 1 illustrate the different steps used for inferring spaces from a long unknown chunk of text using the word cost equation.

Algorithm 1: Infer Spaces (Long_Word)

1: $Words \leftarrow Load\_Dictionary(125k\ Eng)$;
2: $Costs \leftarrow []$;
3: **for** $word\, in\ words$ **do**
4:     $Costs.append(Cost(W))$
5: **end for**
6: $max\_len \leftarrow Max\_length(Words)$;
7: $Detected\_words \leftarrow []$;
8: **for** $i\ in\ range(1,\ length(Long\_Word)$ **do**
9:     $word \leftarrow Best\_match(Word[0:i], Costs, max\_len)$
10:     $Detected\_words.append(word)$;
11: **end for**
12: **return** $Detected\_words$;

The preceding allows us to dynamically turn long and unmatched chunk into appropriate list of tokens that maximize the refeqn:WordCost. For example, the chunck
"iamreallygratefulfortheopportunity".
is turned, after 'Infer-Spaces' algorithm, into:
"i am really grateful for the opportunity".

The overall integration of the 'Infer-Spaces' algorithm in word tokenization and processing pipeline is described in Fig. 3.

In essence, we first apply the standard NLP tokenization module to apply initial preprocessing and generate tokens using the NLTK word tokenizer *word_tokenize* that employs space character as a delimiter for separating individual tokens. Next, we set two thresholds γ_1 and γ_2 on the length of the chunk, which control the discarding and the call to the 'Infer-Spaces' algorithm. This is motivated by the desire to have sufficiently short-length unmatched chunks ignored, and only long-length chunk (beyond γ_2) are inputted to 'Infer-Spaces' algorithm. Second, in order to account for misspelling and typographical errors for unmatched chunk of standard length (between γ_1 and γ_2), an extended dictionary D_{extra} has been created. The latter is constructed in the following way. First, all transcriptions are concatenated to form a transcription corpus. Second, all tokens are matched to the initial 125K dictionary. If a non-match is found, the sentence of the underlined token is manually checked to decide whether a correction can be performed or left as it is. Especially, the abbreviations, youth language and slang words are expanded to their corresponding text, and spelling errors are corrected if they fall within two-fold edit distance. A brief statistics of such construction indicates that unmatched tokens represent around 10% of the total corpus.
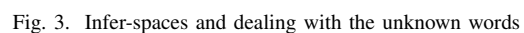
Therefore, D_{extra} is used as a database for correcting all unknown words, assumed misspelling which are either generated by the automatic translation software or user's phonetics ambiguity. In terms of threshold values, unknown tokens whose length is less than **four** characters are removed. On the other hand, words whose length is greater than **ten** characters are systematically passed to the Infer-Spaces algorithm. While unknown words whose length is between **five** and **ten** characters are maintained and passed to the extra dictionary for correction, if any.

Fig. 3. Infer-spaces and dealing with the unknown words

3) Features extraction: We distinguish statistical features and deep-learning related features. The statistical features consist of the following four models:

- Part-of-speech (POS) categories where NLTK POS tagger [36] was used to attach a part of speech tag to each word of the transcription.
- Bag-of-Word model (BOW) with both word and character N-gram level (BOW-Ng) in which we restrict the feature space to to 5000 features [52].
- Term Frequency-Inverse Document Frequency (TF-IDF) model with both word (TF-IDF-Ng) and character (TF-IDF-NgC) N-gram levels [56]. In both BOW-Ng and TF-IDF-NgC. We used all the possible N-grams as one feature space where the N value was in range of 1 and five.
- Emotion category. This is extracted using Text2emotion tool to extract five different emotions (angry, happy, surprise, sad and fear).

Although, the three first aforementioned features were pretty common in a NLP-based approach, the intuition behind the emotion feature arises from the observation that if a person is happy/sad, then this will ultimately impact his/her well-being and possibly his mental health, and thereby, his personality trait. A summary of these features is reported in Table IV:

TABLE IV
STATISTICAL FEATURES

Feature	# Features	N-grams	N-gram-Char
POS	33	-	-
Emotions	5	-	-
BOW	5000	-	-
BOW-Ng	5000	1-5	-
TF-IDF	5000	-	-
TF-IDF-Ng	5000	1-5	-
TF-IDF-NgC	5000	-	3-7

Deep learning related features are primarily dominated by various word-embedding models. Four distinct word embedding models have been employed:

- Word2vec (W2V) [41],
- document2vec (D2V) [31],
- Facebook's FastText [4],
- Glove [45].

Besides, for each of the above models, we considered pretrained (with our dataset) and custom trained models for extracting the embedding features. In essence, we considered three distinct scenarios: customer trained model, pretrained model, and update model. The update model is obtained by augmenting the custom data by our dataset (pretrained). Table V summarizes the employed deep learning features.

---

¹Python package for detecting emotions behind a text

TABLE V
DEEP-LEARNING FEATURES

Feature	# Features	# iterations
W2V Custom	300	1000
Pretrained W2V	300	-
Updated W2V	300	5
D2V	300	1000
Glove	300	1000
FastText Custom	300	1000
Pretrained-FastText	300	-

4) Deep learning model: Figure 4 shows the low-level architecture of our model.

After we tokenise our preprocessed transcriptions, we encode each word by its appropriate index in the vocabulary of the used word embedding model. Next, we pass our encoded tokens to an embedding layer. We initialise the embedding layer's weights with the matrix associated with pretrained word embedding to produce a feature vector. Then we pass the embedded vector to a bidirectional LSTM based on the temporal relationship of the inputs to capture more context from both the right and left sides of the sequence.

More specifically, we followed an embedding layer with an attention layer that enables our model to concentrate on essential pieces of the extracted features to build a context vector c_i from the previous hidden states $s_1...s_N$ and the current h_i ones (Eq.2). Next, the context vector c_i at i^{th} state is computed as the average of the previous states weighted with the attention scores a_i (Eq.4).

$$c_i = \sum_{j=1}^{N} a_{ij} \times s_j \qquad (2)$$

$$a_i = softmax(f_{att}(h_i, s_j)) \qquad (3)$$

The attention function f_{att} calculates an unnormalized alignment score between the current hidden state h_i and the previous hidden state s_j where v_a and W_a are the learned attention parameters.

$$f_{att}(\mathbf{h}_i, \mathbf{s}_j) = \mathbf{v}_a^\top \tanh(\mathbf{W}_a[\mathbf{h}_i; \mathbf{s}_j]) \qquad (4)$$

Next, we attach this attention layer to four sequential fully connected layers (FC) of different sizes. Each layer uses a rectified linear activation function (ReLU) as an activation function with a dropout in the second FC layer to avoid overfitting. Finally, the fourth FC layer is connected to a fully connected layer consisting of one single neuron with a single output which uses a linear function to produce a continuous value between 0 and 1 as a final output.

Nevertheless our initial testing of the above deep learning architecture revealed that such architecture performs well for a single output variable prediction but shows slight degradation of performance in case of multiple variable output estimation. For instance, in the MNIST dataset, it was impossible to predict the colour and the type of cloths with the same branch and get a reasonable accuracy. It sounds in case of multitasking scenarios where independent variables (to be estimated) share the same weight variables, the weights update in each task compromises the network performance. This motivates us to
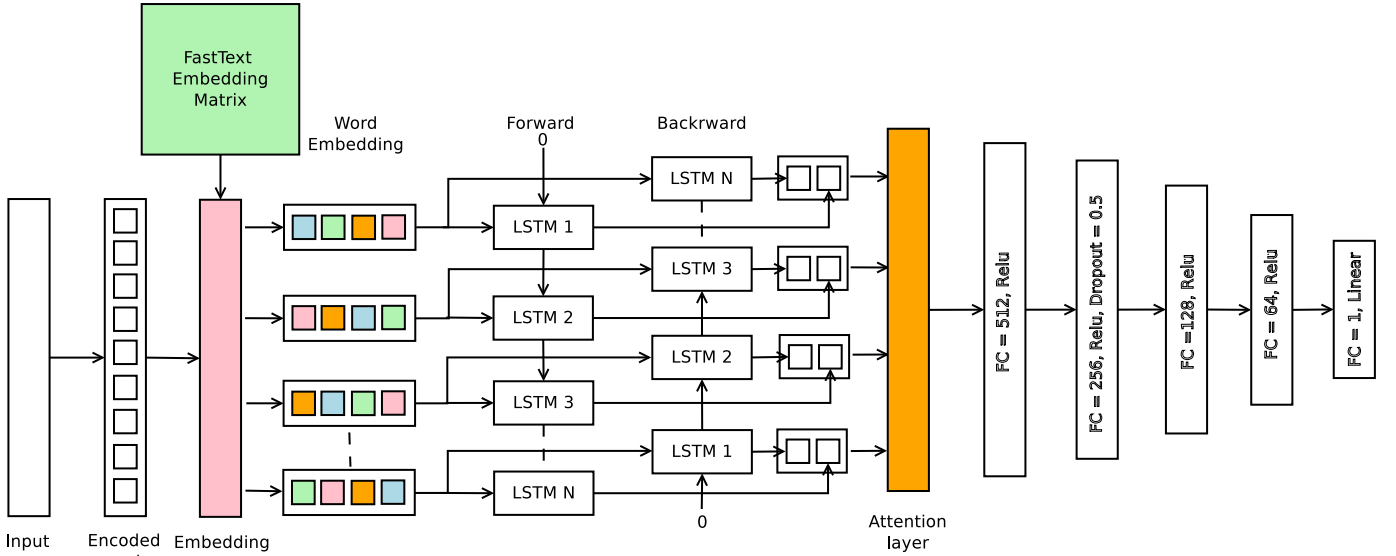
Fig. 4. Deep learning architecture for single output estimation

decouple the estimation of each personality trait and interview score, suggesting a parallel architecture the suggested deep-learning model was used for estimating each variable (personality trait and job interview score) independently of each others. Figure 5 shows our multi-output architecture for Big-Five personality traits and job interview estimation using multiple branches[2].

Implementation details We used the Keras library to implement deep learning models. The training phase varies from one model to another based on the model's depth and the size of the features employed. The training takes around 5 seconds per epoch in which our LSTM model takes 2.6 minutes on an NVIDIA GeForce GTX 1070 GPU for 32 epochs with learning rate equal to 6e-4 and an Adam optimizer.

V. EVALUATION METRICS

For the sake of consistency with the First Impression Challenge tasks, we employed the same performance metric, which consists of the Mean Absolute Error (MAE) defined for each personality trait and job interview score. More specifically, given the prediction $y_p(i)$ and ground truth $x_p(i)$ of pth trait at i^{th} sample, the MAE is given by:

$$E_p = (\frac{1}{n}) \sum_{i=1}^{n} |y_p(i) - x_p(i)| \tag{5}$$

MAE can also be used to compute the accuracy as

$$Accuracy(p) = 1 - E_p \tag{6}$$

On the other hand, by setting up a threshold on the output value, we can turn the prediction of Big Five personality traits into a binary classification (e.g., by setting a threshold value on E_p to 0.5). Therefore, we can estimate the classification accuracy using the Area under the ROC (receiver operating

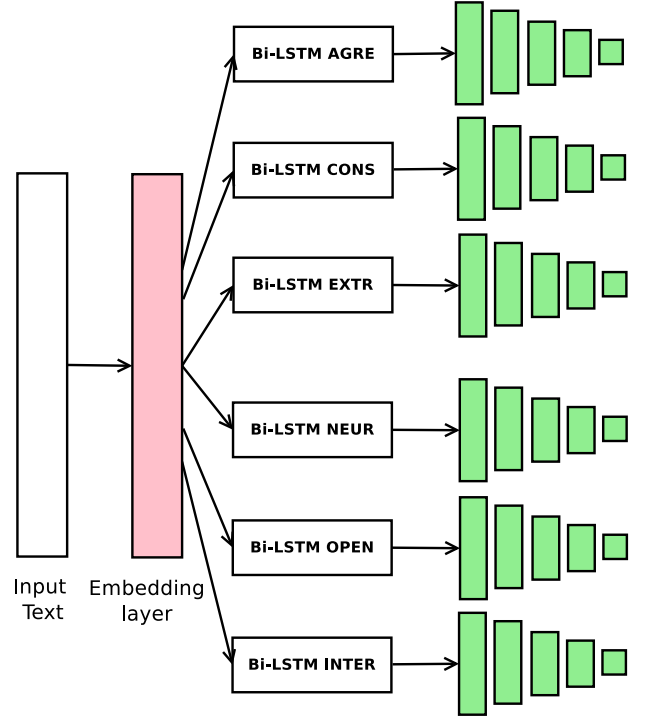[2]https://github.com/bounabyazid/IEEE-trans-Affective-Computing-First-Impression

Fig. 5. Overall architecture for estimation of personality trait and job interview score. The green parts correspond to the deep learning model of a single output starting from the Embedding layer to the output layer

characteristic) curve (AUC). The latter is a graph that shows the performance of a classification model at all classification thresholds where a model with wrong predictions has AUC close to 0. In contrast, a model with correct predictions has an AUC value close to 1.

VI. RESULTS AND DISCUSSIONS

In our experiment, we employed a two-fold strategy where we initially focused to identify the best features and model

parameters that yield the most accurate job interview score with a comparison with other state-of-the-art deep learning models and machine learning regression models. The best model parameters are then replicated in the overall architecture of Figure 5 to predict the six variables (five personality traits and job interview score).

A. single output system examination

In order to test and identify appropriate setting of our single-output deep-learning architecture, we have carried out the task of estimation of job interview score using commonly employed deep-learning features. The latter consists of custom word2vec (w2v), Pretrained-Word2vec (Pre-w2v), Augmented Pretrained and custom Word2vec (Up-Pre-w2v), custom document2vec (D2V), Custom Glove (Glove), Custom FastText (Fast), Pre-trained FastText (Pre-Fast). However, we did not include in the comparison augmented glove, pretrained d2v, and augmented FastText simply because FastText and Glove do not support the continuous training, while there is no pretrained model for D2V.

The results were also compared to seven other deep-learning models, which have reported high score in several text mining classification challenges:

FastText [23], Hierarchical Attention Network (HAN) [63], Region Convolutional Neural Network (R-CNN) and R-CNN variant [30], Text-Bi-LSTM [1], [49], Text-CNN [28], Text-RNN [35].

The results of this investigation are reported in Table VI where the performance of each architecture with a selected feature set is provided. The reading indicates that our sequential model with attention layer was the best performing in overall. Besides, it also shows that the feature set that achieves the best performance is the pre-trained FastText model. The latter is thereby implemented in the overall architecture Figure 5. Although, we do acknowledge that it was a little laborious to decide the best feature representation where the second best feature (Pre-w2v) is only marginally outperformed by FastText by 0.01%.

We also carried out a comparison with machine learning models to see how the deep-learning model differ from popular machine-learning architectures in terms of the estimation of the job interview score. For this purpose, we compared seven commonly employed regression models; namely, Linear Regression, AdaBoost, Support Vector regression, Decision Tree, Random Forest, K-Nearest Neighbor, Gboost and XGboost. While, the set of features investigated consists of: Part-of-Speech (POS), Bag-of-Word (BOW), Bag-of-Words-N-gram (BOW-Ng), TF-IDF, TF-IDF-N-gram, TF-IDF-N-gram character level (TF-IDF-NgC), Emotions. Table VII summarizes the results of the machine learning models. One notices a relatively small variability in performance where the best accuracy (88.92%) is achieved by the Linear Regression classifier when used with POS features. Similarly, one reports good accuracy result in case of emotion feature and a deep drop in classification result between 8.45% and 6.2%. This can be explained by the importance of the part-of-speech tags (usage of verbs, name, adverb, adjective, etc.) and emotion state in

predicting the outcome of the job interview. We also notice that Gboost and XGboost provide relatively high performance scores regardless the feature set employed. Moreover, using the five emotions as features gave promising results using both Gboost and XGboost models. Comparing machine-learning estimation and deep learning estimation also reveals close performance score, with marginal superiority to our deep learning model. Especially, contrary to the machine learning algorithms, there was less variability in model performance for deep-learning models regardless the feature set employed. On the other hand, we shall mention that the parameters of the machine learning algorithms employed in this study have been optimized using the random grid search approach, which justifies to some extent the relatively good performances achieved by these algorithms as well, together with the fact that similar training samples and preprocessing stages were used.

B. Overall estimation and comparison with state-of-the-art architectures

After setting up the structure and inherent parameters of the overall multiple variable output architecture in Figure 5, we test its performance in terms of each individual personality trait and job interview score. For this purpose, we initially report the performance of the classification in terms of the the area under the receiver operating characteristic curve (AUC). In this respect, the results for all human personality traits and job interview value are reported in figure 6. Besides, for a better visualization, we shifted the true positive rate with an offset of 0.05 when we plot our AUC for each personality trait and the job interview value in which the shifting starts from 0 to 0.3. On notices from this plot that the estimation of job interview score (INTER) presents a better agreement with the ideal AUC curve (close to top right and top left), which partly explains the use of Job interview score for tuning the deep learning model parameters, while the agreeableness personality trait presents the less agreement fit with the ideal AUC case.

For illustration purpose, we reported in Table VIII, the best performing methods that used text modality on estimation of Big-Five personality traits and job interview score using ChaLearn LAP'16 dataset, alongside the performance of our method. We notice that our model yields the best performance in job interview score prediction, and achieved either first, second, or third best score in personality trait estimation. Besides, the outperformance, if any, is often quite marginal less than 0.3%, except for openness trait where the outperformance reached 2.4%. We notice that our model marginally underperformed for Agreeableness and Openness traits, possibly because the model was extrapolated from job interview score estimation only. Nevertheless, the model performed better than other approaches in case of Consciousness and Neuroticism. Similarly, our architecture becomes the state-of-art for estimating Job Candidate Screening with a score of 89.10% using text data only.

TABLE VI
DEEP LEARNING ARCHITECTURES USED FOR JOB INTERVIEW ESTIMATION

model/Feature	W2v	Pre-w2v	Up-Pre-w2v	D2V	Glove	Pre-Glove	Fast	Pre-Fast
FastText [23]	88.36	88.87	88.56	88.46	88.80	88.91	88.49	**88.95**
Han [63]	**88.88**	88.89	88.83	**88.83**	88.74	88.41	**88.82**	88.94
RCNN [35]	88.34	88.33	88.23	87.62	87.62	88.49	88.25	87.78
RCNN Variant	88.63	88.75	88.53	88.49	88.24	88.61	88.35	88.80
Text-Bi-LSTM [1]	87.24	88.59	88.48	87.54	88.52	88.5	87.34	87.79
Text-CNN [28]	87.94	88.81	88.28	87.93	88.45	88.34	87.79	88.75
Text-RNN	88.53	88.32	88.76	88.62	**88.86**	88.95	88.23	88.74
Our model	88.26	**89.09**	**88.99**	88.60	88.72	**89.05**	88.24	**89.10**

TABLE VII
MACHINE LEARNING ALGORITHMS USED FOR JOB INTERVIEW ESTIMATION

| model/Feature	POS	BOW	BOW-Ng	TF-IDF	TF-IDF-Ng	TF-IDF-NgC	Emotions
Linear Regression	**88.92**	82.03	81.42	82.12	81.84	83.12	88.30
AdaBoost	88.55	88.37	88.37	88.45	88.44	88.49	88.31
SVR	88.47	88.44	88.37	88.63	88.49	88.63	88.35
D.T	88.50	88.37	88.35	88.48	88.44	88.39	88.30
R.F	88.70	88.70	88.63	88.81	88.73	88.81	87.85
KNN	88.82	87.49	87.51	88.52	88.49	88.53	88.33
Gboost	88.90	**88.80**	**88.82**	88.80	88.70	88.67	**88.40**
XGboost	88.88	88.73	88.68	**88.88**	**88.76**	**88.84**	88.40
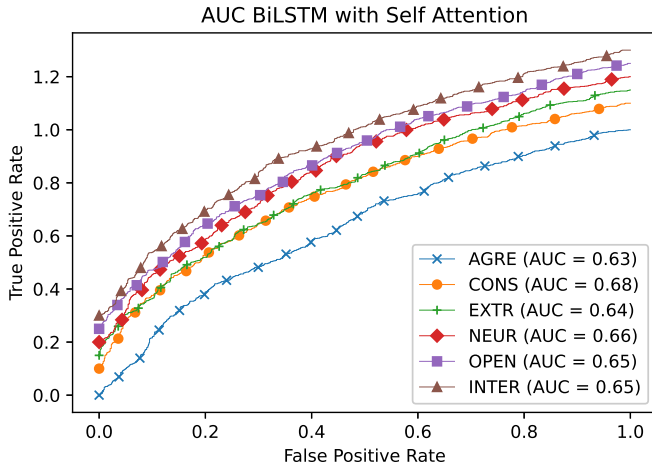
Fig. 6. Area Under the Receiver operating characteristic curve for Human personality and job interview classification

TABLE VIII
A COMPARISON OF THE PROPOSED APPROACH WITH OTHER THAT USE TEXT MODALITY

	AGR	CON	EXT	NEU	OPE	INT
BOW[12]	89.52	87.86	88.15	87.94	88.75	88.45
Skip-Vec[12]	89.71	88.19	88.39	88.27	88.81	88.65
Gorbova et al.[17]	90.10	88.20	87.40	88.40	88.10	88.80
Escalera et al.[62]	89.68	88.00	88.70	88.48	89.03	88.77
Kampman et al. [25]	90.23	87.94	88.23	88.33	91.23	-
Our Method	**89.90**	**88.62**	**88.62**	**88.73**	**89.09**	**89.10**

In overall, given the fact that most of the reported text-modality approaches for estimating personality traits and interview score are rather hybrid-based methods where the text-modality was used in conjunction or in support of either video or audio modality, we consider our solely textual based approach the most performing one. Furthermore, we believe there

is a room for improvement as well by re-visiting the single output fine tuning model by taking into account personality traits together with other feature representation.

Table IX highlights our methods' performance in terms of accuracy for each personality trait and job interview compared to the highest and lowest scores obtained in ChaLearn CVPR 2017 challenge regardless the use of text modality or not. The results show that our method scores are more close to the highest reported scores for the entire challenge than to the lowest ones.

TABLE IX
COMPARISON BETWEEN OUR PROPOSED SYSTEM AND THE LOWEST AND HIGHEST ACCURACY FOR EACH PREDICTION PERSONALITY TRAIT IN THE CHALEARN CVPR 2017 CHALLENGE.

Traits	Our Method	Lowest	Highest
Agreeableness	89,90	89,10	91,37
Conscientiousness	88,62	86,60	91,98
Extraversion	88,62	87,88	92,13
Neuroticism	88,73	86,32	91,46
Openness	89,09	87,48	91,70
Interview	89,10	87,21	92,09

When comparing our results to those who scored high in the competition, it should be pointed out that our findings still are competitive as well. It demonstrates that even if we omit important cues conveyed by video sequence, such as facial traits, gesture and sound variation, we still can achieve an average of 89% in personality traits identification and 89.10% in interview score estimation. The summary also shows a clear difference in terms of minimum and maximum scores reported. It gives a piece of evidence that some of the human personality traits are much more complex to estimate than others. It should also be noted that to promote a fair evaluation of methods for automatically job candidate screening, challenge organizers have only considered personality traits among the seven attributes stated in Mark Cook's book[7] to make interview judgments based only on appearance. Therefore, they focused

on the aspects that are autonomous of the job type to achieve overall results.

VII. Conclusion

This research aims to build a model that predicts and assesses both a job interview and Big-Five personality traits based on video transcriptions. The methodology uses video transcriptions to turn video analytics into text analytics, which allows us to use a well-elaborated NLP pipelines for text analytics. We performed an in-depth analysis using machine learning and deep learning approaches with different text regression models. First, we wanted to identify what is the best model and feature representation for job interview estimation. Then, we used the best model with the best features to construct our deep-learning model. The best model uses bidirectional LSTM with attention layer and the Google pretrained word to vector as the best feature extractor. This model is then replicated in a parallel architecture to estimate both each of the five personality traits and the job interview score. The research has been performed using ChaLearn LAP'16 dataset and achieved a performance of 89% average accuracy in personality trait prediction and 89.10% in job interview assessment score. These results were quite close to the top results achieved in the ChaLearn LAP Challenge. Our results demonstrate the proposal's soundness even when a substantial amount of information is lost since the textual transformation ignores important visual features such as gestures, facial and audio-related cues that cannot be translated to textual data. However, there is always room for further improvement in optimization for both the deep-learning model's architecture and the generation of the appropriate features.

More generally, these findings are consistent with research showing that textual data have the ability to give an insight into a writer's personality. Future research will consider the development of subsequent reasoning to translate both video and audio related patterns to textual data to enhance the accuracy of the prediction models.

References

[1] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[2] S. E. Bekhouche, F. Dornaika, A. Ouafi, and A. Taleb-Ahmed. Personality traits and job candidate screening via analyzing facial videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 10–13, 2017.

[3] J.-I. Biel, L. Teijeiro-Mosquera, and D. Gatica-Perez. FaceTube. In *Proceedings of the 14th ACM international conference on Multimodal interaction - ICMI '12*. ACM Press, 2012.

[4] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

[5] J. N. Butcher. Minnesota multiphasic personality inventory. *The Corsini Encyclopedia of Psychology*, pages 1–3, 2010.

[6] K. Cherry. What are the big 5 personality traits? verywell mind. Retrieved 12 June 2020, from.

[7] M. Cook. *Personnel selection: Adding value through people-A changing picture*. John Wiley & Sons, 2016.

[8] Á. Corral and I. Serra. The brevity law as a scaling law, and a possible origin of zipf's law for word frequencies. *Entropy*, 22(2):224, 2020.

[9] P. T. Costa and R. R. McCrae. Trait theories of personality. In *Advanced personality*, pages 103–121. Springer, 1998.

[10] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor. Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, 18(1):32–80, 2001.

[11] M. Cristani, A. Vinciarelli, C. Segalin, and A. Perina. Unveiling the multimedia unconscious. In *Proceedings of the 21st ACM international conference on Multimedia - MM '13*. ACM Press, 2013.

[12] H. J. Escalante, H. Kaya, A. A. Salah, S. Escalera, Y. Gucluturk, U. Guclu, X. Baró, I. Guyon, J. J. Junior, M. Madadi, et al. Explaining first impressions: modeling, recognizing, and explaining apparent personality from videos. *arXiv preprint arXiv:1802.00745*, 2018.

[13] D. W. Fiske. Consistency of the factorial structures of personality ratings from different sources. *The Journal of Abnormal and Social Psychology*, 44(3):329, 1949.

[14] A. J. Gill and J. Oberlander. Taking care of the linguistic features of extraversion. In *Proceedings of the 24th Annual Conference of the Cognitive Science Society*, pages 363–365, 2002.

[15] L. GOLDBERG, L. Goldberg, L. GOLDBERG, L. Goldberg, L. Goldberg, and R. Goldberg. Language and individual differences: The search for universals in personality lexicons. 1981.

[16] C. Gong, F. Lin, X. Zhou, and X. Lü. Amygdala-inspired affective computing: To realize personalized intracranial emotions with accurately observed external emotions. *China Communications*, 16(8):115–129, 2019.

[17] J. Gorbova, I. Lusi, A. Litvin, and G. Anbarjafari. Automated screening of job candidate based on multimodal video processing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.

[18] Y. Güçlütürk, U. Güçlü, X. Baro, H. J. Escalante, I. Guyon, S. Escalera, M. A. Van Gerven, and R. Van Lier. Multimodal first impression analysis with deep residual networks. *IEEE Transactions on Affective Computing*, 9(3):316–329, 2017.

[19] Y. Güçlütürk, U. Güçlü, M. A. van Gerven, and R. van Lier. Deep impression: Audiovisual deep residual networks for multimodal apparent personality trait recognition. In *European conference on computer vision*, pages 349–358. Springer, 2016.

[20] F. Gürpınar, H. Kaya, and A. A. Salah. Combining deep facial and ambient features for first impression estimation. In *European conference on computer vision*, pages 372–385. Springer, 2016.

[21] F. Gürpınar, H. Kaya, and A. A. Salah. Multimodal fusion of audio, scene, and face features for first impression estimation. In *2016 23rd International conference on pattern recognition (ICPR)*, pages 43–48. IEEE, 2016.

[22] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition, 2015.

[23] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*, 2016.

[24] J. C. S. J. Junior, Y. Güçlütürk, M. Pérez, U. Güçlü, C. Andujar, X. Baró, H. J. Escalante, I. Guyon, M. A. Van Gerven, R. Van Lier, et al. First impressions: A survey on vision-based apparent personality trait analysis. *IEEE Transactions on Affective Computing*, 2019.

[25] O. Kampman, E. J. Barezi, D. Bertero, and P. Fung. Investigating audio, visual, and text fusion methods for end-to-end automatic personality prediction. *arXiv preprint arXiv:1805.00705*, 2018.

[26] S. M. Kassin. *Essentials of psychology*. Prentice Hall, 2003.

[27] H. Kaya, F. Gurpinar, and A. Ali Salah. Multi-modal score fusion and decision trees for explainable automatic job candidate screening from video cvs. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 1–9, 2017.

[28] Y. Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar, Oct. 2014. Association for Computational Linguistics.

[29] R. Kiros, Y. Zhu, R. R. Salakhutdinov, R. Zemel, R. Urtasun, A. Torralba, and S. Fidler. Skip-thought vectors. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[30] S. Lai, L. Xu, K. Liu, and J. Zhao. Recurrent convolutional neural networks for text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

[31] Q. Le and T. Mikolov. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196. PMLR, 2014.

[32] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[33] S. Lestrade. Unzipping zipf's law. *PloS one*, 12(8):e0181987, 2017.

[34] Y. Li, J. Wan, Q. Miao, S. Escalera, H. Fang, H. Chen, X. Qi, and G. Guo. Cr-net: A deep classification-regression network for multimodal apparent personality analysis. *International Journal of Computer Vision*, pages 1–18, 2020.

[35] P. Liu, X. Qiu, and X. Huang. Recurrent neural network for text classification with multi-task learning. *arXiv preprint arXiv:1605.05101*, 2016.

[36] E. Loper and S. Bird. Nltk: The natural language toolkit. *arXiv preprint cs/0205028*, 2002.

[37] J. Mahmud. Ibm watson personality insights: The science behind the service. *IBM Corporation, https://developer. ibm. com/watson/blog/2015/03/23/ibm-watson-personality-insights-science-behind-service*, 2015.

[38] F. Mairesse and M. Walker. Personage: Personality generation for dialogue. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 496–503, 2007.

[39] G. Matthews, I. J. Deary, and M. C. Whiteman. *Personality Traits*. Cambridge University Press, 2009.

[40] R. R. McCrae and P. T. Costa. Validation of the five-factor model of personality across instruments and observers. *Journal of personality and social psychology*, 52(1):81, 1987.

[41] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[42] D. S. Moschona. An affective service based on multi-modal emotion recognition, using eeg enabled emotion tracking and speech emotion recognition. In *2020 IEEE International Conference on Consumer Electronics - Asia (ICCE-Asia)*, pages 1–3, 2020.

[43] W. T. Norman. 2800 personality trait descriptors–normative operating characteristics for a university population. 1967.

[44] J. Pennebaker. Linguistic styles: Language use as an individual difference. . *Journal of personality and social psychology*, 77(6):1296–1312, 1999.

[45] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[46] V. Ponce-López, B. Chen, M. Oliu, C. Corneanu, A. Clapés, I. Guyon, X. Baró, H. J. Escalante, and S. Escalera. ChaLearn LAP 2016: First round challenge on first impressions - dataset and results. In *Lecture Notes in Computer Science*, pages 400–418. Springer International Publishing, 2016.

[47] V. Ponce-Lopez, B. Chen, M. Oliu, C. Corneanu, A. Clapes, I. Guyon, X. Baro, H. J. Escalante, and S. Escalera. Chalearn lap 2016: First round challenge on first impressions-dataset and results. In *European Conference on Computer Vision*, pages 400–418. Springer, 2016.

[48] R. A. Power and M. Pluess. Heritability estimates of the big five personality traits based on common genetic variants. *Translational psychiatry*, 5(7):e604, 2015.

[49] C. Raffel and D. P. Ellis. Feed-forward networks with attention can solve some long-term memory problems. *arXiv preprint arXiv:1512.08756*, 2015.

[50] B. W. Roberts, N. R. Kuncel, R. Shiner, A. Caspi, and L. R. Goldberg. The power of personality: The comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. *Perspectives on Psychological Science*, 2(4):313–345, Dec. 2007.

[51] B. Schuller, S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, R. v. Son, F. Weninger, F. Eyben, T. Bocklet, et al. The interspeech 2012 speaker trait challenge. In *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[52] A. Sethy and B. Ramabhadran. Bag-of-word normalized n-gram models. In *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[53] G. M. Smith. Usefulness of peer ratings of personality in educational research. *Educational and Psychological measurement*, 27(4):967–984, 1967.

[54] C. Stangor, J. Walinga, et al. Introduction to psychology-1st canadian edition. 2018.

[55] A. Subramaniam, V. Patel, A. Mishra, P. Balasubramanian, and A. Mittal. Bi-modal first impressions recognition using temporally ordered deep audio and stochastic visual features. In *Lecture Notes in Computer Science*, pages 337–348. Springer International Publishing, 2016.

[56] M. Umer, I. Ashraf, A. Mehmood, S. Ullah, and G. S. Choi. Predicting numeric ratings for google apps using text features and ensemble learning. *ETRI Journal*, 43(1):95–108, 2021.

[57] C. Ventura, D. Masip, and A. Lapedriza. Interpreting cnn models for apparent personality trait regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.

[58] R. J. W. Vernon, C. A. M. Sutherland, A. W. Young, and T. Hartley. Modeling first impressions from highly variable facial images. *Proceedings of the National Academy of Sciences*, 111(32):E3353–E3361, July 2014.

[59] A. Vinciarelli and G. Mohammadi. A survey of personality computing. *IEEE Transactions on Affective Computing*, 5(3):273–291, 2014.

[60] R. Wang, S. Huang, Y. Zhou, and Z. G. Cai. Chinese character handwriting: A large-scale behavioral study and a database. *Behavior Research Methods*, 52(1):82–96, 2020.

[61] X.-S. Wei, J.-H. Luo, J. Wu, and Z.-H. Zhou. Selective convolutional descriptor aggregation for fine-grained image retrieval, 2016.

[62] A. S. Wicaksana and C. C. Liem. Human-explainable features for job candidate screening prediction. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1664–1669. IEEE, 2017.

[63] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, San Diego, California, June 2016. Association for Computational Linguistics.

[64] C.-L. Zhang, H. Zhang, X.-S. Wei, and J. Wu. Deep bimodal regression for apparent personality analysis. In *European conference on computer vision*, pages 311–324. Springer, 2016.

[65] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27, 2015.

Yazid Bounab Ph.D. condidate University of Oulu, Finland. He is currently working as a doctoral researcher in CMVS Research Group at the University of Oulu, Finland, His research interests include Natural language processing, machine learning, pattern recognition, image processing,

Mourad Oussalah . He is currently working as an associate professor in CMVS Research Group at the University of Oulu, Finland. His research interests include computer vision, image processing, pattern recognition, machine learning, deep learning, natural language processing, and social network analysis.

Nabil Arhab Ph.D. condidate University of Oulu, Finland. He is currently working as a doctoral researcher in CMVS Research Group at the University of Oulu, Finland, France. His research interests include Natural language processing, machine vision, deep learning, image processing,

Salah Eddine Bekhouche received his Ph.D. degree in electronics from the University of Biskra, Algeria in 2017. He is currently a postdoctoral researcher in EPAN Research Group at the University of Technology of Belfort-Montbéliard (UTBM), France. His research interests include computer vision, pattern recognition, and image processing.