EXPLORING SOFT SKILLS INDICATORS IN MULTIPLE MINI INTERVIEWS (MMI) SCENARIO RESPONSES

Ryan Huynh[1], Lee Gillam[2] and Alison Callwood[3]

[1]School of Computer Science, University of Surrey,
Guildford, United Kingdom
2Sammi-Select Ltd, Guildford, United Kingdom
3School of Health Sciences, University of Surrey,
Guildford, United Kingdom

ABSTRACT

Multiple mini-interviews (MMIs) are a widely used and validated interview method for eliciting soft skills. By using multiple, separate, and timed interviews in which each has a distinct scenario, MMIs purportedly reduce possibilities such as a biased individual dictating results, although potentially inconsistent scoring by interviewers may still impact on fairness. However, MMIs overall can be seen as challenging to run due to the number of interviewers and assessments required. In this paper, we discuss the progress in automatically, and consistently, extracting soft skills from transcriptions of MMI responses to support such assessment. While previous research has focused on extracting soft skills from job postings and written responses, to the best of our knowledge there is no other published research on soft skill extraction from MMI responses. We begin by annotating collected MMI responses to assure presence of soft skills, then evaluate the effectiveness of combining word embeddings with classifiers to identify soft skill indicators. The most promising result, F1-Score = 0.79, compares favourably to previous literature on extracting soft skills from other datasets and is encouraging of further exploration.

KEYWORDS

Soft Skills Extraction, Multiple Mini Interviews, MMI, Word2Vec, BERT

1. INTRODUCTION

Interviews, of various kinds and constructions, are a common part of recruitment. In a competitive job market, identification of a suitable candidate may entail multiple rounds of interviews and other assessments. Such processes may raise concerns regarding fairness and the equitable evaluation of each candidate if, for example, a lone, vocal, assessor could dictate the entire outcome.

Despite the potential for biases in evaluation, the typical objective of such interviews is to evaluate the backgrounds and skills of each candidate to determine their suitability to the desired role. These skills can be categorised into two groups: hard skills and soft skills. Hard skills relate to job-specific proficiencies, such as programming experience for a software engineering role [1]. Soft skills encompass more fundamental competencies, including effective communication, teamwork, and leadership [2]. Together, both types of skills provide valuable information for evaluation.

David C. Wyld et al. (Eds): IoTE, CNDC, DSA, AIAA, NLPTA, DPPR - 2024
pp. 49-63, 2024. - CS & IT - CSCP 2024 DOI: 10.5121/csit.2024.142305

Existing research on the evaluation of hard and soft skills in interviews widely acknowledges that soft skills are inherently subjective and more challenging to quantify [3-6]. All of these studies stated that hard skills can easily be evaluated through tests and assignments. Whereas, soft skills are more difficult due to the lack of clear and consistent standards, methods, and criteria. This subjectivity is further compounded when soft skills are assessed by multiple interviewers, each with their own expectations and preferences. Thus, a framework is sought for the consistent evaluation of soft skills.

Multiple mini-interviews (MMIs) are a widely used and validated interview method for eliciting soft skills. MMIs involve multiple, separate, timed interviews based on scenarios designed to target and assess specific soft skills within interviewee responses [7]. However, concerns remain about, for example, consistency of scoring when interviewers are conducting the same MMI multiple times throughout an entire day [8] and whether this and other potential interviewer biases could impact on fairness and, ultimately, selection [1,9,10].

To address consistency, this paper explores whether indicators of soft skills can be extracted from MMI responses. The research question guiding this study is: Can machine learning methods be employed to automatically and consistently identify indicators of soft skills in MMI scenario responses? Our goal is to develop a system that aligns closely with human judgement so as to inform and/or support it – though, consistent with emergent legislation, not to replace it. We note that such a system would itself have to be evaluated for bias, as well as the importance of producing explainable results. However, such work is beyond the scope of the present paper.

Methodologically, our exploration uses a dataset that was previously collected from past MMIs, and a stage of manual annotation ensures the presence of indicators of soft skills. Word embeddings, first from Word2Vec so as to offer methodological comparability to prior works, and then from BERT, are used to generate sentence embeddings. Multi-class classification using these embeddings then supports evaluation.

It is important to clarify that this study is a small-scale, exploratory experiment aimed at demonstrating the potential of soft skill extraction methods in MMI responses. The limited extent of training data used in this study means that while it provides valuable insights, it is not intended to offer a definitive conclusion. The research serves as a preliminary step to justify further investigation and development.

The contributions of this paper are as follows:

1) We assess the feasibility of using machine learning to automatically identify soft skill indicators within MMI transcriptions, addressing a gap in the literature where most research has focused on job advertisements and CVs.
2) We propose a novel application of identifying soft skills within MMI transcriptions using word embeddings, specifically Word2Vec and BERT. This extends the use of these techniques from their traditional applications to a new domain, providing methodological comparability and advancing the field.
3) We provide an analysis of the performance of different embedding methods and classifiers in the context of skill extraction.

Although our focus is on MMIs, we believe that the methodologies employed here are highly adaptable for the development of recommender systems tailored to the HR sector. For instance, the techniques used in our study to extract soft skills from textual data can potentially be applied to build a recommender system that matches candidate's soft skills with job description. This is supported by research such as that by Gugnani and Misra [11] who presented a job recommender

system that matches resumes to job descriptions by extracting implicit skills. These parallels suggest that our approach could be transferred into HR applications, particularly in the automation and improvement of candidate-job matching processes.

The remainder of the paper is structured as follows: Section 2 provides a background overview, Section 3 discusses related work of soft skills extraction, as seems primarily directed towards job advertisements and CV's. Section 4 outlines the approach, and Section 5 presents our results. Section 6 summarises findings and limitations, and offers indications of possible future directions.

2. BACKGROUND

Use of multiple mini-interviews (MMIs) has mainly been demonstrated in medical and health science recruitment as a more efficacious method of assessing candidates compared to traditional interviews, due to various factors such as effectiveness, acceptability, fairness, and exposure to different interviewers [12-15]. The aim is to reduce interviewer bias and provide a more comprehensive evaluation of each candidate's soft skills. MMIs involve a succession of structured interviews designed to evaluate both cognitive attributes, such as critical thinking, creativity and reasoning, and non-cognitive attributes, such as communication, leadership, and teamwork skills [7]. Cognitive attributes can be related to hard skills and may be considered more measurable [16]; non-cognitive attributes and soft skills are often characterised by their less tangible nature, pertaining to personal qualities and behaviours that are less readily quantified [17].

In 2004, Eva et al [7] first reported the application of MMIs, which are organised as a timed circuit of stations. Candidates can begin the process at any station within the circuit, and at each station, they will participate in a short-timed interview. Each mini-interview usually involves an openended, scenario-based question to which interviewees respond, with a bare minimum of prompting by the interviewer. On station completion, candidates rotate to the next station until all stations have been attended. Each station is staffed with a minimum of one designated interviewer who poses exactly the same question to every candidate in the process. In principle, this should ensure consistency at each station, in contrast to having different interviewers, although potential for inconsistency does exist – for example due to interviewer mood or fatigue [18], or with biases such as the halo effect [19] where interviewers might also be tempted to offer beneficial prompts to some candidates but not others. Assessment of each candidate's response by the interviewer at each station, in a brief gap between interviews, typically involves scoring using a Likert scale, although alternative methods such as global rating [20,21] and checklist scoring [22,23] have also been used. Candidates are evaluated by collating scores across all interviewers. While summation of scores from all stations remains the most prevalent method for determining a candidate's final score, a modified borderline approach [24] has also been applied in some contexts.

The onset of SARS-CoV-2 forced many institutions using MMIs to alter their approaches or adopt various virtual multiple mini-interviews (VMMIs): numerous organisations conducted MMIs through video platforms like Zoom [25-29] with widely varying degrees of fidelity to conduct in the physical setting or typical operational practices. Use of VMMIs may, however, be more costeffective than MMIs as evidenced by Tiller et al. [30], who demonstrated that the absence of need of a venue and travel can result in cost savings of up to six times compared to traditional in-person MMIs.

A high-fidelity VMMI approach that incorporates pre-recorded questions, and runs as a timed circuit, has been reported recently [31]. Interviewees are offered a time period (in days) within which to be assessed, at their convenience, and respond to the questions online as a form of

digital interview. Once they choose to start, the same time constraints of a physical MMI are applied and interviewees are presented with their time-constrained scenario questions. Responses are recorded (video and audio) for each question, with the short break remaining in place between questions to mimic the gap being used for assessment but also for the candidate to gather their thoughts. With pre-recorded questions, every candidate receives the same question delivery, removing any concerns as to interviewer mood or fatigue at this point, and, in contrast to the physical setting, all candidates start the circuit from the same station. Interviewers can assess the recordings at convenient, perhaps less fatigued, times, although other biases and inconsistent scoring across candidates could not yet be ruled out. Additionally, responses can be transcribed to offer a rich source of text data that, allied to interviewer scores for attributes, provides for exploration of indicators of soft skills and how they are being quantified.

3. LITERATURE REVIEW

Given transcribed MMI responses, it becomes possible to consider automation of skill extraction. Indicatively, there is almost twice as much literature of potential relevance to extraction of hard skills than to extraction of soft skills - Google Scholar suggests some 616,000 results for [extraction of hard skills] vs 314,000 for [extraction of soft skills], likely with an extent of overlap. The majority of these papers discuss analysing standard recruitment documents such as job postings or advertisements, and responses in CV's and cover letters – and for recommendation, matching between recruitment documents and responses. To the best of our knowledge, there is presently no literature that addresses analysis of MMI responses. We found approaches explained in Khaouja et al. [32], Fareri et al. [33] and Sayfullina et al. [34] are, methodologically, of interest.

Khaouja et al. [32], presented a methodology for constructing a soft skill taxonomy. They compiled a dataset of 380,000 English job advertisements, 20,000 from various careers websites and 360,000 from a Kaggle dataset (https://www.kaggle.com/c/job-salary-prediction/data). They extracted all bi and tri-grams phrases which had a frequency of at least 10. To form the initial set of soft skills, all extracted bi- and tri-grams were validated on DBpedia, a structured knowledge base derived from Wikipedia. If a webpage existed, that phrase is considered as a skill. Manual verification was conducted on this set resulting in 24 soft skills extracted. Since this initial set was small, they added soft skills found in works of two other studies [35,36], a further 26 soft skills were added resulting in a total set of 50.

To gather alternative labels for each soft skill within this set, they queried both DBpedia and Word2Vec. Using DBpedia, for each soft skill, all internal direct and backwards hyperlinks were extracted. Whereas, for Word2vec, they used the jobs advertisements to train the Word2Vec model and extracted related terms for each soft skill. All terms which were common between the sets from DBpedia and Word2vec were then used to form a final set of related words for each soft skill.

For evaluation, they acquired and manually annotated random samples of job advertisements. They used a soft skill tagging method where they utilised the soft skills and their related words to match the appearance of these soft skills in the job advertisements, compared to the manual annotations. They calculated true positives (correct skills extracted from job ad), false positives (skills extracted that do not exist in job ad) and false negatives (skills not extracted although they exist in job ad) and measured precision and recall, yielding an F1-score of 0.84.

Fareri et al. [33] created SkillNER, a named entity recognition (NER) system, to automatically extract fragments of text that mention a soft skill. To build this system, they gathered and formed a dataset containing the abstracts of 5,000 scientific papers. The papers selected for the dataset

contained the phrase “soft skills’’ in the title, abstract or keywords. A group of experts annotated each abstract by sentence and extracted fragments that mention a soft skill. The extracted fragments contain two components: the entity (the skill itself), and the clue (a set of terms, lexical expressions, or recurrent patterns associated with the appearance of the soft skill). In their paper, only the top 10 soft skills were presented.

Using this data, they trained and compared two types of classification models. One approach used feature-based supervised learning with a support vector machine (SVM), while the other employed deep learning through a multilayer perceptron (MLP). For encoding the text into numerical vectors that these models could process, they utilised the Word2Vec method.

The evaluation is similar to Khaouja et al. [32] assessing true positives, false positives, and false negatives, for previously unseen job advertisements, though a different set of advertisements to those of Khaouja et al. [32]. Results show that the feature-based approach, SVM, achieved the higher F1-score of 0.73, whereas the deep learning approach of MLP achieved an F1-score of 0.62. Whilst they are using the SVM for SkillNER, the authors pose the potential for neural networks, and in particular Bidirectional Encoder Representations from Transformers (BERT) by Devlin et al. [37] for soft skill extraction, but deem it beyond the scope of that work.

Sayfullina et al. [34] present a phrase-matching-based approach for detecting soft skills in unstructured text. They classified this problem as a binary text classification problem where the prediction is made if the soft skill is present or not. The dataset used is created through a combination of two datasets, one comprising job advertisements (the dataset as used by Khaouja et al. [32]) and CVs (manually collected 525 CV from indeed.com). The corpus was manually annotated to label whether a sentence contained a soft skill, each sentence was verified by two additional annotators. This process led to two classes, positive (soft skill exists in the sentence) and negative (soft skill does not exist in the sentence). These were split into a training set and two testing sets (Job test and CV test), where the training and Job test sets consisted of sentences from the job ads dataset, and the CV test set only consisted of sentences from the CV dataset.

All words in sentences were embedded using Word2Vec, and experimented with three neural network classifiers: convolutional neural networks (CNN), long short-term memory (LSTM), and hierarchical attention networks (HAN) for binary classification. Results on the job test set demonstrated LSTM achieved the highest performance, followed by CNN and HAN, with F1scores of 0.84, 0.81, and 0.80, respectively. Whereas, the results on the CV test set demonstrated HAN achieved the highest performance, followed by LSTM and CNN, with F1-scores of 0.90, 0.81, and 0.80, respectively.

Khaouja et al. [32], Fareri et al. [33] and Sayfullina et al. [34] inform our methodology. In contrast to their uses of recruitment documents such as job advertisements or responses such as CVs, our dataset comprises MMI responses. This dataset encompasses both the candidates' responses and the specific scenarios designed to evaluate various soft skills. Manual annotation will similarly be applied to support evaluation. To encode words, Word2Vec and BERT encoder will be used. Several models, including transformers, will be tested. For evaluation, and to indicate comparability, we will similarly report F1-score and comprised measures where relevant. At this stage, we seek only to determine if such an approach could become useful for extraction; scoring responses, as MMI interviewers would, is beyond the scope of the present paper.

4. METHODOLOGY

Khaouja et al. [32], Fareri et al. [33] and Sayfullina et al. [34] use and/or produce lists of soft skills, and their experiments inherently relate to some kind of labelling of appearance of items in this list, or of words/phrases with similar meaning. For instance, we applied the soft skills tagging approach from Khaouja et al. [32] and SkillNER from Fareri et al. [33] to our dataset to determine if any soft skills could be detected. Our findings revealed that only a limited number of soft skills were identified: understand, understands, kind, compassion, kindness, decisions, and positive. This indicates that such lists would only be suitable if MMI responses included explicit soft skill words/phrases like "teamwork," "communication," or "conflict management". However, MMI responses are expected to provide evidence of soft skills in use rather than mere mention. There is a need, then, to determine words/phrases that are indicators for the soft skills of interest (we may refer to the soft skills as attributes), to be able to relate a potential multiplicity of these in any given sentence or response.

4.1. Data

To the best of our knowledge, there is no publicly available data on MMI or VMMI scenario responses. Therefore, a dataset was used that had previously been collected from a VMMI approach for student admissions to some undergraduate degree programs; note that this data cannot be publicly shared due to legal and ethical reasons. Interviewees participated in an online assessment featuring multiple scenario-based questions, with their responses transcribed for subsequent analysis. The total data obtained from 2409 candidates comprise responses across 7 scenario questions (non-uniform), totalling 8,142,679 words (tokens).

For these experiments, we decided to focus on one scenario. This scenario expects an interviewee to describe how they would react in a social setting with a friend who abruptly departed in anger, triggered by the recent loss of a parent. It is designed to assess three scenario-specific attributes: (i) Communication; (ii) Compassion/Empathy; and (iii) Respect/Dignity. From our corpus of data, a subset of ten responses to this scenario was randomly selected. Table 1 shows the sentence and word counts for each response.

Table 1. Sentence and Word Counts for 10 Responses – 197 sentences comprising 3891 words

Response	Sentence Count	Word Count
1	23	453
2	17	531
3	13	346
4	36	536
5	16	343
6	25	311
7	8	404
8	25	245
9	15	316
10	19	406

4.1. Manual Annotation

Let X represent the set of all responses across all candidates. Here, $x \in X$denotes the set of responses for each candidate, and consequently, $y \in x$ represents a response to a particular question or scenario for a specific candidate. With i^{th} input, we can denote $\alpha_y{}^i = [j_1, j_2, \ldots, j_n]$as the representation of each sentence within a response, while $\beta_y{}^i = [k_1, k_2, \ldots, k_n]$ signifies the scores assigned to each attribute for the sentence. Our objective is to predict $\beta_y{}^i$ based on $\alpha_y{}^i$. At this stage, ahead of determining how appearance relates to scoring, we substitute binary digits denoting appearance for scores by manually annotating each sentence.
To generate $\alpha_y{}^i$ each response was manually split into individual sentences. To create feature set $\beta_y{}^i$ we followed the manual annotation approach of Sayfullina et al. [34], examining the dataset at sentence level and labelling indicators of soft skills, which was guided by an expert MMI interviewer. Here, we can distinguish a fragment as a word/phrase that is an indicator of one or more soft skills, as will be beneficial for verification, and relate a vector of three binary digits where each represents appearance or otherwise of indicators for attributes.

Consider, for example, a response sentence “I would approach the young man” - “approach” is identified as an indicator of Communication so is the relevant fragment for this sentence, and for attributes in the order Communication, Compassion/Empathy, and Respect/Dignity, the resulting vector for this sentence is [1, 0, 0].

Similarly, the fragment “calmly and kindly inform” would be associated with both Communication and Compassion/Empathy - “inform” relates Communication, while “calmly and kindly” relates Compassion/Empathy - for the sentence “I would calmly and kindly inform the young man” producing [1, 1, 0]. Table 2 shows the prevalence of soft skill indicators in the annotated dataset.

Table 2. Data Distribution for each Attribute

Attribute	Sentence Count	Word Count	Average Number of Words per Sentence
Communication	58	1837	32
Compassion/Empathy	71	2349	33
Respect/Dignity	41	1323	32
None	82	778	9

4.1. Word Embeddings

For compatibility of approach with Khaouja et al. [32], we first explore how word embeddings from Word2Vec might benefit the approach. Neural networks necessitate numerical representations of words through vectorisation, where input words tend to be transformed into a numerical vector with respect to positions in a large vocabulary. A set of resulting (numeric) associations somewhere between the inputs and outputs is often referred to as the word embeddings, and as well being used to indicate outputs related to specific inputs these embeddings have been used for various purposes and in various ways. Prominent vectorisation methods include term frequency-inverse document frequency (TF-IDF) [38], global vectors for word representations (GloVe) [39] and, as we use here initially, Word2Vec [40].

Google's pre-trained Word2Vec model relates 3 million words (vocabulary/types) through embeddings. This model was trained on approximately 100 billion words (tokens), sourced from the Google News dataset. During training, two algorithms were employed: Continuous Bag of

Words (CBOW) and Skip-gram. CBOW offers likely words for a specified context, whereas Skipgram offers context for specific words. Each word will be converted into a numerical vector embedding composed of 300 dimensions, a value established to be optimal by Google.

Returning to our recent example, we can obtain embeddings for each word in the sentence "I would approach the young man" and use these to produce a whole-sentence embedding through vector addition. This way, each sentence has a 300-dimension representation. While averaging word embeddings is a more common approach, we chose vector addition because of how the Word2Vec creators reported it being useful in preserving word ordering within sentences [41]
Google's pre-trained Word2Vec model can provide sentence embeddings for our case. However, we can further optimise these embeddings by using transfer learning. Transfer learning is a technique that fine-tunes a model on a corpus of data to boost its performance on the target task. By providing more related data, we can adjust embeddings for specific words, hoping to generate better representations. As a result, we can use the majority of our authentic data to further train Google's Word2Vec model in hopes of identifying any differences in performance.

Continuing our investigation of embeddings, and as suggested by Fareri et al. [33] Bidirectional Encoder Representations, or BERT from Transformers, offers an alternate route for encoding semantic information. BERT, developed by Devlin et al. [37] includes a bidirectional training approach that allows the model to consider context from both preceding and following words at the same time. This bi-directionality improves comprehension of word meanings in context by capturing complex relationships inside sentences. In contrast to Word2Vec's fixed-size embeddings, BERT embeddings have a dimensionality of 768, which is much greater than Word2Vec's 300 dimensions. Contextual embeddings for each word in a sentence are generated and aggregated to produce a full sentence embedding, resulting in a more fine-grained representation of the input texts.

4.2. Modelling

As mentioned previously, each sentence is represented by a vector of three binary numbers representing indicators of specific soft skills. These vectors are transformed into class 'IDs', forming a total of 8, and hence creating a multi-class classification problem. These class 'IDs', along with embeddings, are presented to classifiers in order to associate soft skills. We use the scikit-learn (Python) package for this, as well as for the train_test_split function, and then 'Lazy Predict'(https://lazypredict.readthedocs.io/en/latest/index.html), a publicly available tool to apply more than 25 classification models from which we can capture metrics such as model accuracy, balanced accuracy, and for comparison against literature cited previously, F1-score.

4.3. Evaluation

We aimed to adopt the approach outlined by Sayfullina et al. [34] where they employed two test sets. For the first test set (T1), it consists of sentences split from the dataset in a 70/30 ratio for training/testing purposes. However, since we intend to evaluate the models using complete responses, we gathered three random additional responses for the second testing set (T2). These responses had sentence counts of 24, 26 and 11, and word counts of 603, 577 and 404 respectively. All were manually annotated, and class predictions are obtained using Lazy Predict.

5. RESULTS

5.1. Experiment I - Word2Vec

Table 3 presents the results of the top 6 classifiers.

Table 3. Data Distribution for each Attribute

Model	T1 Acc.	T2 Acc.	T1 F1	T2 F1
Perceptron	0.57	0.61	0.53	0.62
Nearest Centroid	0.50	0.65	0.47	0.64
LGBM	0.47	0.60	0.42	0.54
Logistic Regression	0.53	0.64	0.47	0.63
SGD	0.53	0.61	0.46	0.60
GaussianNB	0.47	0.56	0.43	0.62

Nearest Centroid classifier boasts the highest accuracy and F1-score on the T2 dataset (0.65 and 0.64 respectively). This classifier, which identifies a class centroid during training and measures unseen data points' similarity to these centroids, suggests that the data points are well separated by their labels and that the classifier can capture the essential characteristics of each class.

The Perceptron model is considered the building block for deep learning, known as the first and simplest neural network and therefore reportedly proficient in binary classification tasks [42]. However, its performance may vary depending on the characteristics of the data. In this case, the Perceptron model achieves the highest accuracy and F1-score on the T1 dataset (0.57 and 0.53 respectively), indicating that it can learn the linearly separable patterns in the data well. However, on the T2 dataset, its accuracy and F1-score drop to 0.61 and 0.62 respectively, which are lower than those of Nearest Centroid. This suggests that the Perceptron model may not be able to generalise well to new or unseen data, as it may overfit to the training set and fail to capture the underlying patterns of the T2 test set. A possible solution to improve the performance of the Perceptron model on the T2 dataset could be to apply regularisation techniques, such as L2 norm or dropout. L2 norm is a technique that adds a penalty term to the loss function, proportional to the squared magnitude of the weights. This prevents the weights from becoming too large and reduces the complexity and variance of the model. Even though the Perceptron model is the basic unit of artificial neural networks, its results are good in comparison to the other classifiers, which show promise for exploring other deep learning models.

5.2. Experiment II - Transfer Learning

In this experiment, we aimed to explore the impact of using transfer learning on word embeddings.

Specifically, we fine-tuned Google's pre-trained Word2Vec model to see how it affects performance. For this scenario, we had a total of 933 responses in our dataset, of which we used 75%: 700 responses, yielding 18251 sentences. We used Lazy Predict to produce results from the same classifiers, see Table 4 for the results.

Table 4. Results using Transfer Learning to Fine-tune Word2vec

Model	T1 Acc.	T2 Acc.	T1 F1	T2 F1
Perceptron	0.79	0.61	0.79	0.61
Nearest Centroid	0.76	0.67	0.74	0.67
LGBM	0.80	0.55	0.76	0.52
Logistic Regression	0.79	0.54	0.78	0.54
SGD	0.74	0.62	0.73	0.62
GaussianNB	0.62	0.54	0.70	0.48

The use of transfer learning does seem to have a positive effect. This is mainly shown by the T1 test set, which has increased average accuracy and F1-scores in all classifiers. The biggest difference was within the Perceptron model. Comparing the average accuracy and F1-score to Table 3 we see that the accuracy increased by 0.22 (from 0.57 to 0.79) and the F1-score increased by 0.26 (from 0.53 to 0.79). It seems like fine-tuning the Word2Vec model to adjust word embeddings does have a significant impact on boosting performance on the T1 test set. However, for the Perceptron model, results from the T2 test showed similar performance as in Table 3, where no fine-tuning was conducted. Therefore, further experiments with an increased test size are needed to determine whether transfer learning has an effect.

Looking at the results from the T2 test set, the Nearest Centroid classifier had the highest average accuracy and F1-score of 0.67, an increase of 0.02 and 0.03 respectively, compared to Table 3. We observed that this classifier performed relatively well across both T2 testing accuracy and F1scores for the attributes Communication and Respect/Dignity, but it showed a decrease of approximately 0.15 when it came to the attribute Compassion/Empathy. Looking further, we identified a significant decrease between the training and testing metrics for this attribute, which suggests overfitting. When referring back to table 2 we noticed that this attribute had the largest amount of training sentences (71).

The application of transfer learning, in our case, updating the embeddings of Google's Word2Vec model, produced mixed results, in which metrics on the T1 test set showed significant increase but did not on the T2 test set.

5.3. Experiment III - BERT Embeddings

Here we replaced Word2Vec with BERT to identify any performance differences using the same classifiers. In-addition, with the Perceptron model showing promise within deep learning, and whilst utilising BERT embedding, it made sense to experiment using a pre-trained BERT model (https://huggingface.co/bert-base-uncased). We used this setup: a learning rate of 2e-5, a batch size of 16, and 10 epochs. Table 5 presents the results.

Table 5. Results using BERT Embeddings

Model	T1 Acc.	T2 Acc.	T1 F1	T2 F1
Perceptron	0.77	0.68	0.78	0.64
Nearest Centroid	0.73	0.65	0.73	0.64
LGBM	0.77	0.63	0.73	0.53
Logistic Regression	0.79	0.65	0.78	0.60
SGD	0.77	0.65	0.76	0.56
GaussianNB	0.72	0.56	0.72	0.40
BERT	0.84	0.71	0.55	0.62

Comparing results of these to Table 4, except for Perceptron, all the other classifiers showed similar or slightly lower accuracy and F1-score with BERT embeddings than with Word2Vec embeddings. This could be explained by the larger number of features in BERT (768) than in Word2Vec (300). Nearest Centroid and Perceptron had the same T2 F1-score of 0.64, but Perceptron had a slightly higher T2 accuracy of 0.68. Perceptron's performance might stem from its neural network architecture, which can handle more complex and high-dimensional data better than the other classifiers.

The pre-trained BERT model performs reasonably well, achieving the highest T1 and T2 testing accuracy compared to all previous experimentations with 0.84 and 0.71, respectively. However, its T2 F1-score remains slightly lower than that of Nearest Centroid. Nonetheless, there is promise in improving the F1-score through various optimisation techniques. For instance, fine-tuning the model by adding more layers can help capture complex patterns and relationships in the data that were previously overlooked. Altering weighting schemes can also address class imbalances, ensuring that each class contributes proportionally to the final F1-score. In future iterations, experimenting with different architectures and loss functions could increase the F1-score. This does suggest that deep learning models seem to be in the direction of interest.

5.4. Summary

Results indicate that the Perceptron delivers the best performance in our experiments. Additionally, deep learning models, particularly when using BERT embeddings, also demonstrated potential. We believe that further fine-tuning and increasing the volume of training data could enhance the performance of these deep learning models. Moreover, consistency under both increases of data volume and number of soft skills (attributes) needs to be investigated, and it is important to monitor how classifier results vary as such additions are made.

6. CONCLUSIONS,LIMITATIONS AND FUTURE WORK

This paper has discussed challenges posed by MMIs and our explorations of automated soft skill extraction as may support alleviation of some of these. Our exploration involved several key steps: using a dataset of MMI responses, annotating these responses for soft skills, deriving embeddings from the data using Word2Vec and BERT, and applying a range of classifiers to evaluate the soft skill labels. The most notable outcome, achieved through transfer learning with Word2Vec and a Perceptron model (F1-Score = 0.79) appears to be comparable to approaches in literature that have involved extracting soft skills from other datasets and provides a suitably encouraging base from which to explore further.

Although our study primarily focused on MMIs, we acknowledge the potential to extend our findings to the development of recommender systems in the HR sector. This study's methods for soft skill extraction could be adapted to build a recommender system tailored for HR professionals, particularly in candidate-job matching scenarios. In an MMI setup, where a set of stations assess candidates for a single job or role, each candidate is evaluated against attributes such as communication, empathy, or teamwork. These attribute scores could be used not only to rank candidates for a specific role but also to compare them across different roles. In a multiple candidates vs. multiple jobs setting, as is typical in recommender systems, explicit mention of soft skills in job descriptions could be matched with the extracted soft skill indicators from candidates' responses.

It is vital to understand that there are several limitations of this study. We readily acknowledge that the extent of training data used within this paper is limited, and it would be desirable to work with much more of the data if the annotation workload can be addressed. This paper presents a small-scale, exploratory experiment intended to establish the potential of soft skill extraction methods in MMI responses and serves as a justification for further work. Similar exploratory studies have laid the foundation for larger research efforts [43,44], and we believe this work will do the same. It is, however, important to evaluate whether an approach has suitable potential prior to such investment of such effort, and the data brings two further challenges to bear: (i) the number of soft skills to address is higher; (ii) MMIs involve scoring (Likert scale) whilst the present work, and related past work, focuses only on existence.

Methodologically, there is a wealth of possibility – use of Word2Vec, initially to make for comparability to the closest-related prior work, is limiting when other approaches have been shown to be better for a variety of natural language tasks (e.g. SentenceBERT [45] for embeddings), and results from BERT here suggest that such past work might also be merit reappraisal.
It is important to acknowledge, also, that both Word2Vec and BERT embeddings, as revealed by prior research [46-50], are susceptible to biases, particularly but not only gender biases. However, the evaluation of these biases is beyond the scope of this study. Although bias in machine learning models is a critical concern, particularly in systems handling sensitive data like MMIs, addressing bias will require further, dedicated research. Concerns about bias should also be explored in the data used, for example in the absence of data at the lowest or highest scores - a kind of response bias in scoring. Selection of classifiers may change according to changes in earlier selections, and the need for explanations of results, or use of these in examination of consistency of human judgement, also remains to be addressed. The latter would also relate to a limitation on how to use such a system, as certain countries and entities, including the EU, seek to legislate [51] to prevent life-changing decisions from being made by automated approaches.

REFERENCES

[1] M. K. P. So, A. M. Y. Chu, and A. Tiwari, “Interviewer bias when using multiple mini-interviews in selecting student nurses in a Chinese setting,” *Nurse Education Today*, vol. 121, p. 105676, Feb. 2023. doi:10.1016/j.nedt.2022.105676

[2] J. J. Heckman and T. Kautz, “Hard evidence on Soft Skills,” *Labour Economics*, vol. 19, no. 4, pp. 451–464, Aug. 2012. doi:10.1016/j.labeco.2012.05.014

[3] A. Zhang, “Peer assessment of soft skills and hard skills,” *Journal of Information Technology Education: Research*, vol. 11, pp. 155–168, 2012. doi:10.28945/1634

[4] G. Zheng, C. Zhang, and L. Li, “Practicing and evaluating soft skills in it capstone projects,” *Proceedings of the 16th Annual Conference on Information Technology Education*, vol. 23, pp. 109–113, Sep. 2015. doi:10.1145/2808006.2808041

[5] G. I. Continisio *et al.*, "Evaluation of soft skills among Italian healthcare rehabilitators: A Cross Sectional Study," *Journal of Public Health Research*, vol. 10, no. 3, Jun. 2021. doi:10.4081/jphr.2021.2002

[6] D. (Tres) Bishop, "The Hard Truth About Soft Skills," *Muma Business Review*, vol. 1, pp. 233–239, Dec. 2017. doi:10.28945/3803

[7] K. W. Eva, J. Rosenfeld, H. I. Reiter, and G. R. Norman, "An admissions OSCE: The multiple mini-interview," *Medical Education*, vol. 38, no. 3, pp. 314–326, Mar. 2004. doi:10.1046/j.13652923.2004.01776.x

[8] I. van der Spuy, A. Busch, and J. Bidonde, "Interviewers' experiences with two multiple miniinterview scoring methods used for admission to a master of physical therapy programme," *Physiotherapy Canada*, vol. 68, no. 2, pp. 179–185, May 2016. doi:10.3138/ptc.2015-24e

[9] M. Ross *et al.*, "Are female applicants rated higher than males on the multiple mini-interview? findings from the University of Calgary," *Academic Medicine*, vol. 92, no. 6, pp. 841–846, Jun. 2017. doi:10.1097/acm.0000000000001466

[10] I. W. Incoll, J. Atkin, J. R. Frank, S. Vrancic, and O. Khorshid, "Gender Associations with selection into Australian Orthopaedic Surgical Training: 2007–2019," *ANZ Journal of Surgery*, vol. 91, no. 12, pp. 2757–2766, Nov. 2021. doi:10.1111/ans.17320

[11] A. Gugnani and H. Misra, "Implicit skills extraction using document embedding and its use in job recommendation," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 08, pp. 13286–13293, Apr. 2020. doi:10.1609/aaai.v34i08.7038

[12] F. Patterson *et al.*, "How effective are selection methods in medical education? A systematic review," *Medical Education*, vol. 50, no. 1, pp. 36–60, Dec. 2015. doi:10.1111/medu.12817

[13] S. Razack *et al.*, "Multiple mini-interviews versus traditional interviews: Stakeholder acceptability comparison," *Medical Education*, vol. 43, no. 10, pp. 993–1000, Oct. 2009. doi:10.1111/j.13652923.2009.03447.x

[14] S. Uijtdehaage, L. "Hy Doyle, and N. Parker, "Enhancing the reliability of the multiple miniinterview for selecting prospective health care leaders," *Academic Medicine*, vol. 86, no. 8, pp. 1032–1039, Aug. 2011. doi:10.1097/acm.0b013e3182223ab7

[15] J. R. Clark, C. A. Miller, and E. L. Garwood, "Rethinking the admissions interview: Piloting multiple mini-interviews in a graduate psychology program," *Psychological Reports*, vol. 123, no. 5, pp. 1869–1886, Dec. 2019. doi:10.1177/0033294119896062

[16] J. Lamri and T. Lubart, "Reconciling hard skills and soft skills in a common framework: The Generic Skills Component Approach," *Journal of Intelligence*, vol. 11, no. 6, p. 107, Jun. 2023. doi:10.3390/jintelligence11060107

[17] J. E. Humphries and F. Kosse, "On the interpretation of non-cognitive skills: What is being measured and why it matters," *SSRN Electronic Journal*, pp. 174–185, Apr. 2016. doi:10.2139/ssrn.2879804

[18] S. Humphrey, S. Dowson, D. Wall, V. Diwakar, and H. M. Goodyear, "Multiple mini-interviews: Opinions of candidates and interviewers," *Medical Education*, vol. 42, no. 2, pp. 207–213, Jan. 2008. doi:10.1111/j.1365-2923.2007.02972.x

[19] C. Towaij, N. Gawad, K. Alibhai, D. Doan, and I. Raîche, "Trust me, I know them: Assessing interpersonal bias in surgery residency interviews," *Journal of Graduate Medical Education*, vol. 14, no. 3, pp. 289–294, Jun. 2022. doi:10.4300/jgme-d-21-00882.1

[20] C. A. Terregino, M. McConnell, and H. I. Reiter, "The effect of differential weighting of academics, experiences, and competencies measured by multiple mini interview (MMI) on Race and ethnicity of cohorts accepted to one medical school," *Academic Medicine*, vol. 90, no. 12, pp. 1651–1657, Dec. 2015. doi:10.1097/acm.0000000000000960

[21] L. R. Hopson *et al.*, "The multiple mini-interview for emergency medicine resident selection," *The Journal of Emergency Medicine*, vol. 46, no. 4, pp. 537–543, Apr. 2014. doi:10.1016/j.jemermed.2013.08.119

[22] C. Roberts, N. Zoanetti, and I. Rothnie, "Validating a multiple mini-interview question bank assessing entry-level reasoning skills in candidates for graduate-entry medicine and dentistry programmes," *Medical Education*, vol. 43, no. 4, pp. 350–359, Apr. 2009. doi:10.1111/j.13652923.2009.03292.x

[23] S. Kolagari, Z. Sabzi, M. Modanloo, and N. Behnampour, "The validity and reliability of the multiple mini-interview in assessing the capabilities of Nursing Education phd candidates: A methodological study," *Bangladesh Journal of Medical Science*, vol. 21, no. 4, pp. 788–794, Sep. 2022. doi:10.3329/bjms.v21i4.60249

[24] K. W. Eva *et al.*, "Predictive validity of the multiple mini-interview for selecting medical trainees," *Medical Education*, vol. 43, no. 8, pp. 767–775, Aug. 2009. doi:10.1111/j.13652923.2009.03407.x

[25] K. D. Inzana, R. Vanderstichel, and S. J. Newman, "Virtual multiple mini-interviews for veterinary admissions," *Journal of Veterinary Medical Education*, vol. 49, no. 3, pp. 273–279, Jun. 2022. doi:10.3138/jvme-2020-0107

[26] N. Singh, C. DeMesa, S. Pritzlaff, M. Jung, and C. Green, "Implementation of virtual multiple mini-interviews for fellowship recruitment," *Pain Medicine*, pp. 1717–1721, Apr. 2021. doi:10.1093/pm/pnab141

[27] T. Ungtrakul, W. Lamlertthon, B. Boonchoo, and C. Auewarakul, "Virtual multiple Miniinterview during the Covid-19 pandemic," *Medical Education*, vol. 54, no. 8, pp. 764–765, Jun. 2020. doi:10.1111/medu.14207

[28] S. Yolanda, W. Wisnu, J. M. Wahjudi, and A. Findyartini, "Adaptation of internet-based multiple mini-interviews in a limited-resource medical school during the coronavirus disease 2019 pandemic," *Korean Journal of Medical Education*, vol. 32, no. 4, pp. 281–289, Dec. 2020. doi:10.3946/kjme.2020.175

[29] V. Sabesan *et al.*, "Implementation and evaluation of virtual multiple mini interviews as a selection tool for entry into Paediatric Postgraduate Training: A Queensland experience," *Medical Teacher*, vol. 44, no. 1, pp. 87–94, Aug. 2021. doi:10.1080/0142159x.2021.1967906

[30] D. Tiller *et al.*, "Internet-based multiple mini-interviews for candidate selection for graduate entry programmes," *Medical Education*, vol. 47, no. 8, pp. 801–810, Jul. 2013. doi:10.1111/medu.12224

[31] A. Callwood *et al.*, "Feasibility of an automated interview grounded in multiple Mini interview (MMI) methodology for selection into the Health Professions: An International Multimethod Evaluation," *BMJ Open*, vol. 12, no. 2, Feb. 2022. doi:10.1136/bmjopen-2021-050394

[32] I. Khaouja, G. Mezzour, K. M. Carley, and I. Kassou, "Building a soft skill taxonomy from job openings," *Social Network Analysis and Mining*, vol. 9, no. 1, Aug. 2019. doi:10.1007/s13278019-0583-9

[33] S. Fareri, N. Melluso, F. Chiarello, and G. Fantoni, "Skillner: Mining and mapping soft skills from any text," *Expert Systems with Applications*, vol. 184, p. 115544, Dec. 2021. doi:10.1016/j.eswa.2021.115544

[34] L. Sayfullina, E. Malmi, and J. Kannala, "Learning representations for soft skill matching," *Lecture Notes in Computer Science*, pp. 141–152, Jul. 2018. doi:10.1007/978-3-030-11027-7_15

[35] M. Daneva, C. Wang, and P. Hoener, "What the job market wants from requirements engineers? an empirical analysis of online job ads from the Netherlands," *2017 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*, Nov. 2017. doi:10.1109/esem.2017.60

[36] L. K. Yanaze and R. de Deus Lopes, "Transversal Competencies of Electrical and Computing Engineers considering market demand," *2014 IEEE Frontiers in Education Conference (FIE) Proceedings*, pp. 1–4, Oct. 2014. doi:10.1109/fie.2014.7044169

[37] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, *Proceedings of the 2019 Conference of the North*, Jun. 2019. doi:10.18653/v1/n19-1423

[38] S. Robertson, "Understanding inverse document frequency: On theoretical arguments for IDF," *Journal of Documentation*, vol. 60, no. 5, pp. 503–520, Oct. 2004. doi:10.1108/00220410410560582

[39] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation,"

[40] *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, Oct. 2014. doi:10.3115/v1/d14-1162

[41] T. Mikolov *et al.*, "Efficient Estimation of Word Representations in Vector Space," arXiv preprint arXiv:1301.3781, Jan. 2013.

[42] T. Mikolov *et al.*, "Distributed Representations of Words and Phrases and their Compositionality," Advances in neural information processing systems, vol. 26, Oct. 2013.

[43] D. Lopez-Bernal, D. Balderas, P. Ponce, and A. Molina, "Education 4.0: Teaching the basics of KNN, Lda and simple perceptron algorithms for binary classification problems," *Future Internet*, vol. 13, no. 8, p. 193, Jul. 2021. doi:10.3390/fi13080193

[44] F. Calanca, L. Sayfullina, L. Minkus, C. Wagner, and E. Malmi, “Responsible team players wanted: An analysis of soft skill requirements in job advertisements,” *EPJ Data Science*, vol. 8, no. 1, Apr. 2019. doi:10.1140/epjds/s13688-019-0190-z

[45] C. Si, D. Yang, and T. Hashimoto, “Can llms generate novel research ideas? A large-scale human study with 100+ NLP researchers,” arXiv.org, https://arxiv.org/abs/2409.04109

[46] N. Reimers and I. Gurevych, “Sentence-bert: Sentence embeddings using Siamese BertNetworks,” *Proceedings of the 2019 Conference on Empirical Methods in Natural Language*

[47] *Processing and the 9th International Joint Conference on Natural Language Processing (EMNLPIJCNLP)*, pp. 3982–3992, Aug. 2019. doi:10.18653/v1/d19-1410

[48] M. Babaeianjelodar, S. Lorenz, J. Gordon, J. Matthews, and E. Freitag, “Quantifying gender bias in different corpora,” *Companion Proceedings of the Web Conference 2020*, Apr. 2020. doi:10.1145/3366424.3383559

[49] I. Garrido-Muñoz , A. Montejo-Ráez , F. Martínez-Santiago , and L. A. Ureña-López , “A survey on bias in DEEP NLP,” *Applied Sciences*, vol. 11, no. 7, p. 3184, Apr. 2021. doi:10.3390/app11073184

[50] S. Jentzsch and C. Turan, “Gender bias in Bert - measuring and analysing biases through sentiment rating in a realistic downstream classification task,” *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pp. 184–199, Jul. 2022. doi:10.18653/v1/2022.gebnlp-1.20

[51] D. Petreski and I. C. Hashim, “Word embeddings are biased. but whose bias are they reflecting?,” *AI & SOCIETY*, vol. 38, no. 2, pp. 975–982, May 2022. doi:10.1007/s00146-022-01443-w

[52] T. Leteno, A. Gourru, C. Laclau, and C. Gravier, “An investigation of structures responsible for gender bias in Bert and distilbert,” *Lecture Notes in Computer Science*, pp. 249–261, Apr. 2023. doi:10.1007/978-3-031-30047-9_20

[53] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 on Artificial Intelligence (AI Act), Official Journal of the European Union, vol. OJ L 195, 2024. Available:https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689