Automatic identification of storytelling responses to past-behavior interview questions via machine learning

Adrian Bangerter, Eric Mayor, Skanda Muralidhar, Emmanuelle Kleinlogel, Daniel Gatica-Perez, Marianne Schmid Mast

Received: 28 April 2022 | Revised: 17 March 2023 | Accepted: 19 March 2023

DOI: 10.1111/ijsa.12428

RESEARCH ARTICLE

International Journal of Selection and Assessment | WILEY

Automatic identification of storytelling responses to past-behavior interview questions via machine learning

Adrian Bangerter[1] | Eric Mayor[2] | Skanda Muralidhar[3] | Emmanuelle P. Kleinlogel[4] | Daniel Gatica-Perez[3] | Marianne Schmid Mast[5]

[1]Institute of Work and Organizational Psychology, University of Neuchâtel, Neuchatel, Switzerland

[2]Department of Clinical Psychology and Epidemiology, University of Basel, Basel, Switzerland

[3]Idiap Research Institute, Martigny, Switzerland

[4]Centre d'Economie et de Management de l'Océan Indien, University of Reunion Island, Saint-Denis, France

[5]Department of Organizational Behavior, University of Lausanne, Lausanne, Switzerland

Correspondence
Adrian Bangerter, Institute of Work and Organizational Psychology, University of Neuchâtel, Rue Emile-Argand 11, 2000 Neuchâtel, Switzerland.
Email: adrian.bangerter@unine.ch

Funding information
Schweizerischer Nationalfonds zur Förderung der Wissenschaftlichen Forschung

Abstract

Structured interviews often feature past-behavior questions, where applicants are asked to tell a story about past work experience. Applicants often experience difficulties producing such stories. Automatic analyses of applicant behavior in responding to past-behavior questions may constitute a basis for delivering feedback and thus helping them improve their performance. We used machine learning algorithms to predict storytelling in transcribed speech of participants responding to past-behavior questions in a simulated selection interview. Responses were coded as to whether they featured a story or not. For each story, utterances were also manually coded as to whether they described the situation, the task/action performed, or results obtained. The algorithms predicted whether a response features a story or not (best accuracy: 78%), as well as the count of situation, task/action, and response utterances. These findings contribute to better automatic identification of verbal responses to past-behavior questions and may support automatic provision of feedback to applicants about their interview performance.

KEYWORDS

machine learning, past-behavior question, selection interview, storytelling

Practitioner points

- Past-behavior questions constitute a best practice in selection interviews.
- Past-behavior questions invite applicants to tell a story about what they did in a past work-related situation.
- Applicants often fail to produce stories, and when they do, they tend to focus on describing the situation rather than what they did and what results they obtained.
- Coaching may help them improve their responses but is costly.
- Using machine learning, we accurately predict storytelling responses to past-behavior questions and their narrative content from transcripts of applicant responses.
- It is feasible to design systems for automatic delivery of feedback to applicants to improve their responses to past-behavior questions.

1 | INTRODUCTION

Structured behavioral interviews constitute the state-of-the-art in selection interview practice. They evidence high predictive validity and are gaining traction in many organizations worldwide (Kantrowitz et al., 2018; Levashina et al., 2014; Motowidlo et al., 1992; Roulin, 2017; Roulin et al., 2012; Turner, 2004). Structured interviews often feature past-behavior questions, where applicants are asked to describe their actions in a past work situation. Past-behavior questions constitute invitations to the applicant to tell a story (Ralston et al., 2003). Recruiters score their responses using rating scales to investigate skills, abilities, and previous performance of applicants. However, applicants have difficulties producing good stories on demand, often resorting to pseudostories, that is, abstract, generic descriptions of situations. When they do tell stories, they tend to describe situations in some detail, but often neglect to describe tasks, actions, and results obtained (Bangerter et al., 2014). The fact that applicants differ widely in the extent to which they produce good stories in response to past-behavior questions may limit the validity of structured interviews. It is therefore important to help applicants improve their question-answering behavior in order for recruiters to be able to appropriately evaluate applicants' characteristics. Coaching-based interventions involving delivery of tailored feedback to applicants based on manual analyses of their prior question-answering performance (Maurer & Solamon, 2006; Maurer et al., 2008; Ralston et al., 2003) have been developed, but they are costly.

The current study uses machine learning algorithms to automatically identify whether a response to a past-behavior question constitutes a story or not, and if so, what narrative elements (discourse about situation, task, action, or results) it contains. Our findings can be used to evaluate applicants' responses to past-behavior questions or to generate automatic feedback to help train applicants to provide better answers.

2 | BEHAVIORAL INTERVIEWING AND STORYTELLING RESPONSES TO PAST-BEHAVIOR QUESTIONS

Interviews are among the most widely used selection tools to measure a range of constructs (Huffcutt et al., 2001). Their predictive validity varies as a function of the interview structure. Structured interviews can attain high levels of validity (Huffcutt & Arthur, 1994; Wiesner & Cronshaw, 1988). The behavioral interview is a kind of structured interview centered around the use of behavioral questions. These can be either *past-behavior questions* or *situational questions* (Motowidlo, 1999). Both involve asking applicants questions about their actions in work situations. Past-behavior questions ask applicants to describe what they did in a work situation in the past. An example is *Can you tell me about an occasion where you had to put aside your own work to help a colleague?* Situational questions ask applicants to imagine a fictitious work situation and describe what they would do in it (e.g., *Imagine you received a phone call from a client who has a question about their contract. What do you do in such a situation?*). Applicants' answers are then rated using behaviorally anchored rating scales, which provide raters with definitions and examples of different response levels (e.g., poor, average, or good responses). There is ample evidence for the validity of applicants' responses to behavioral questions in predicting future work performance (Taylor & Small, 2002).

Past-behavior questions constitute invitations to applicants to tell a story. Good stories will reveal information about the typical behavior (Klehe & Latham, 2006), personality, or values (Ralston et al., 2003) of the narrator. They constitute a primary means by which applicants engage in impression management (Stevens & Kristof, 1995). Good storytelling performances may index desirable personal characteristics like charisma or leadership (Sharma & Grant, 2011). At a minimum, a story should proceed from a description of the initial situation to the actions of the applicant and the results obtained. Indeed, a mnemonic device used by recruiters and applicants to organize or evaluate stories is "STAR" (situation, task, action, result; Kessler, 2006).

Good storytelling is a complex task (Tross & Maurer, 2008). Applicants do not always produce stories in response to past-behavior questions. They often resort to pseudostories, generic and abstract descriptions of typical work situations. When they do produce stories focused on a specific episode, they tend to focus on describing the initial situation but neglect to describe their actions in detail, or the results attained by those actions (Bangerter et al., 2014). This may be due to the difficulties in finding a maximally relevant example to narrate under the stressful conditions of the interview (Brosy et al., 2016; Huffcutt et al., 2017). Providing applicants with information about upcoming past-behavior questions does not seem to help them, but interviewer probing improves story production and fosters a more balanced combination of narrative elements (STAR) in applicants' stories (Brosy et al., 2020).

Coaching-based interventions may also help improve applicants' response behavior (Maurer & Solamon, 2006; Maurer et al., 2008). A training program for past-behavior questions showing applicants how to use the STAR method improved their performance (Tross & Maurer, 2008), as did a program featuring image-based narrative intervention (Lukacik et al., 2022). Such programs are based on a mixture of lectures, discussions, exercises, role-play, practice, and feedback. Their drawback is that they are costly and time-consuming, often requiring the services of an expert coach or counselor. Thus, there is potential for using automatic analyses of applicant responses as a basis for delivering feedback as part of a training program, to save time or to partly or fully automate the coaching process.

3 | MACHINE LEARNING ANALYSES OF APPLICANT BEHAVIOR IN INTERVIEWS

The advent of audio- and video-recorded interviews, for example, in the format of aysnchronous video interviews (AVIs; Lukacik et al., 2022) opens new avenues for the automatic analysis of applicant

behavior, especially coupled with machine learning (Liem et al., 2018). Videorecorded selection interviews provide a rich set of applicant data for potential analysis: Head and upper-body video of applicants as well as a speech signal that can be processed for prosodic cues or recognition of words, thereby paving the way for various kinds of content analysis. Machine learning analyses of applicant behaviors typically fall in the domain of supervised machine learning, where the goal is to find a function relating inputs (or observations, or *features*) to outputs (or predicted outcomes, or *labels*) in a way that is both effective and that generalizes to other data (Liem et al., 2018). Machine learning proceeds from a training phase to a test phase using various algorithms (e.g., support vector machines, decision trees, or neural networks).

Several studies have investigated the potential of machine learning for supporting hiring decisions, by predicting outcomes like hireability or personality from applicants' responses to past-behavior questions in either simulated or real interviews (e.g., Biel et al., 2013; Chen et al., 2017; Hickman, Bosch, et al., 2022; Muralidhar et al., 2016; Naim et al., 2018; Nguyen et al., 2014; Rupasinghe et al., 2016; Suen et al., 2020). Nguyen et al. (2014) video-recorded mock face-to-face interviews with a human interviewer and applicant. They predicted hireability from a combination of applicant visual (e.g., nodding, smiling) and audio features (speaking activity, prosody) as well as relational cues (combinations of behavior from both applicant and interviewer). Hickman, Bosch, et al. (2022) predicted personality from a combination of linguistic inquiry and word count (LIWC) variables, *n*-grams, and paraverbal and nonverbal behaviors.

Studies have also developed algorithm-based systems automatically analyzing behavioral parameters to train interviewees (e.g., Gebhard et al., 2019; Heimerl et al., 2022; Hoque et al., 2013; Langer et al., 2016; Lin-Stephens et al., 2022). Langer et al. (2016) automatically recorded interviewees' nonverbal behavior, which was then fed back to them in terms of behavioral recommendations. The intervention helped interviewees reduce interview anxiety and improved performance and perceived hireability. Heimerl et al. (2022) used a combination of automatically extracted facial and body behavioral features and conversational cues (turn-taking computed from vocal cues) to generate textual feedback via a generative adversarial network (GAN)-based approach. Few studies have focused on feedback related to how applicants should modulate what they say, that is, their verbal behavior, even though the effect of verbal behavior on employment decisions is at least as important as those of nonverbal or paraverbal behavior (Burnett & Motowidlo, 1998; Hollandsworth et al., 1979; Rasmussen, 1984). No study we are aware of has tried to automatically analyze storytelling responses to past behavior questions for purposes of designing tailored coaching feedback.

4 | THIS STUDY

Machine learning analyses of applicant responses to past-behavior questions could help applicants, for example, by detecting (in) appropriate responses to provide feedback and improve future performance. The current study intends to further link behavioral interviewing and machine learning interview research. The key element that is missing is the identification of appropriate verbal responses to past-behavior questions, especially storytelling. The goal of this study is thus to compare a suite of algorithms to detect the presence or absence of storytelling and their narrative elements from transcribed speech.

The setup of the study and its different steps are depicted in Figure 1. Data for this study is drawn from an experimental job interview simulation. Participants played the role of applicants, answering four past-behavior questions each. Responses were videorecorded and transcribed, and transcripts were preprocessed into a format enabling automatic content analysis. We manually coded two sets of labels (see Figure 1), that we tried to predict using maching learning. First, each response was coded as to whether it featured a story or not. Second, each story was decomposed into utterances to assess its narrative elements. Each utterance was manually coded as to whether it described the situation, the task or an action undertaken by the participant, or results obtained. This yielded a score (number of utterances) for situation, task/action, and result per story told (i.e., each of these scores constitutes a label we attempted to predict).

We then automatically computed features to predict the labels (see Figure 1). This was done in two ways. First, we used the LIWC package (Pennebaker et al., 2015) and its dictionaries to build features. Second, we computed TF-IDF (term frequency-inverse document frequency) scores, a measure of a term's relative importance in a corpus, as a further set of features (Ramos, 2003).

Finally, we applied different machine learning algorithms to predict the labels (story/not story; counts of STAR) from the features (see Figure 1). Predictions of storytelling (story/not story) relied on support vector machines and random forest algorithms, and predictions of STAR counts relied on random forests, bagged classification and regression trees (CART), and partial least squares (PLS).

5 | METHOD

5.1 | Participants

There were 102 French-speaking participants (58% women, M_{age} = 29.1, SD = 9.5). Half were students and were recruited on campus. The other half were professionals and were recruited by email. Students had <5 years of professional experience and professionals had >5 years of professional experience. Participants held higher-education degrees in fields covering the humanities and human sciences or were studying toward such degrees.

5.2 | Procedure

Participants were recruited for a mock job interview in French in a laboratory setting. Recruitment incentives included a combination of

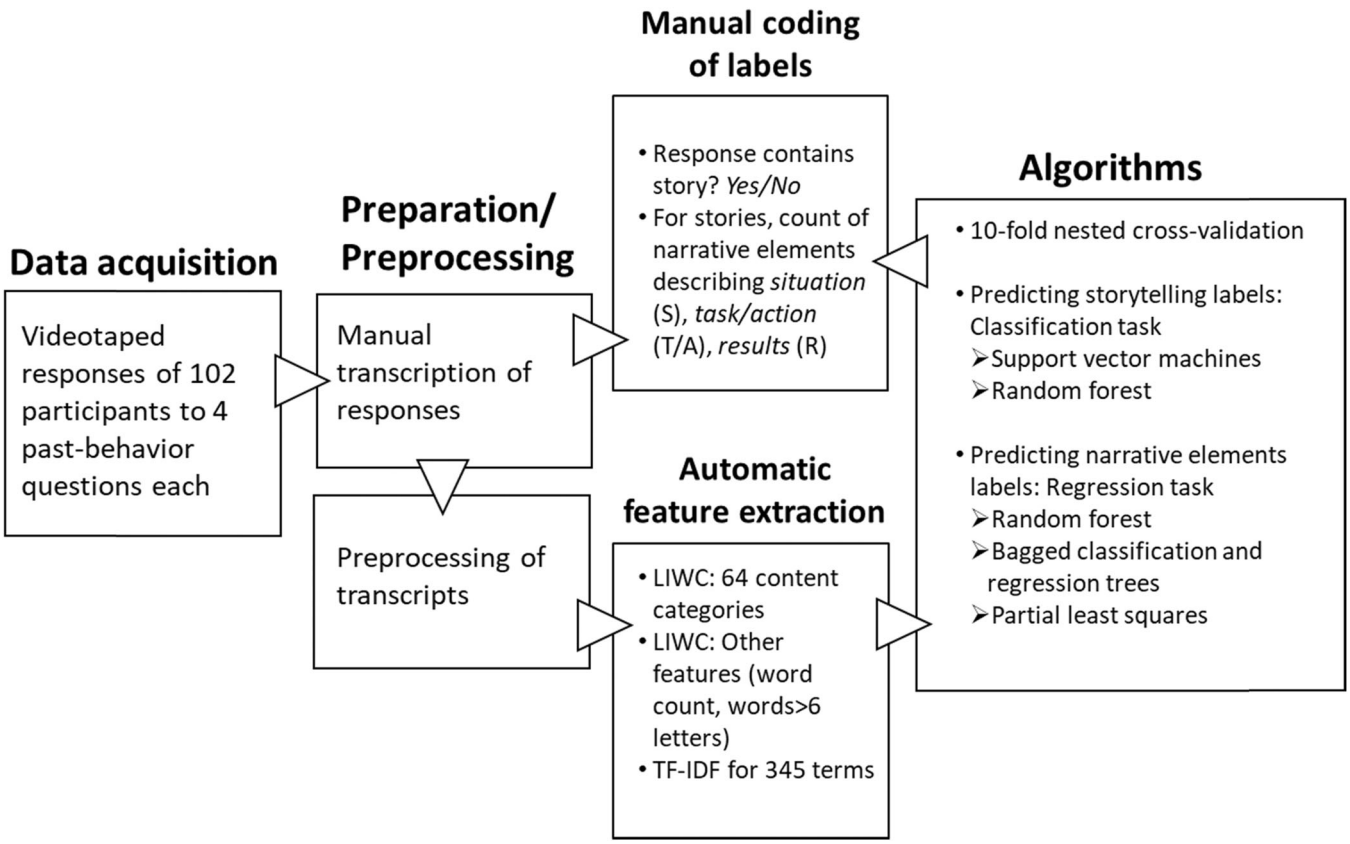

FIGURE 1 Setup and different steps of the study.

personal coaching and monetary incentives. Participants completed a consent form and an online questionnaire and reviewed a job advertisement for a fictitious position corresponding to their sector of activity. About 1 week later, they came to the lab for the interview. They were briefed about the study by a trained experimenter, who also played the role of interviewer. They then were handed the job advertisement again, and had a few minutes to prepare for the interview. The interview consisted of two introductory self-presentation questions (*Could you present yourself in a few words? What are your skills and knowledge that could be of interest to us?*), followed by four past-behavior questions: (1) *Tell me about a situation in which you had to participate in a project with people whose ideas differed from yours*, (2) *Describe to me a situation where you took an initiative that you managed to bring to completion*, (3) *Can you tell me about a situation in which you had to manage several tasks in parallel?*, and (4) *Give me an example of an unexpected situation you had to deal with that forced you to reorganize work already planned*. Interviewers listened and took notes during the interview, produced backchannels (e.g., *mhm*, *okay*), but did not ask probe questions or provide further information, unless prompted by the participants' requests for clarification. The questions were selected because of the relevance of the assessed competencies to all activity sectors targeted by the job advertisements prepared for the study. Interviews were videorecorded (with frontal or 45° offset view of the participants). After the interview, participants filled in additional questionnaires and forms, were debriefed and paid, and received the promised coaching tips.

5.3 | Data preparation

5.3.1 | Transcription and preprocessing

Responses were originally transcribed verbatim from the video files, including hesitations, repetitions (e.g., "pre-pre-presentation") and truncated words as well as comments, which were indicated between parentheses. The transcripts were then preprocessed in R. This included the removal of comments, that is, content between parentheses and the parentheses, replacement of dashes in truncated words with spaces, replacement of multiple spaces by one space and the homogenization of hesitations to a form recognized by LIWC in French (Piolat et al., 2011). We used a French-language stemming procedure (Benesty, 2019; based on Savoy, 1999) to reduce the vocabulary size of the TF-IDF features and thus dimensionality of the feature set (Hickman, Thapa, et al., 2022). There were 345 TF-IDF features after stemming. Because stemming may potentially collapse valid differences between words, we did not use stemming on the LIWC features.

5.3.2 | Manual coding of labels: Storytelling and STAR

Each response to each question (mean word count 201.6, median = 171, *SD* = 123.1) was coded for the presence or absence of a story by a main coder (the data set consists of four responses per participant, that is, 408 responses). A story was defined as a series of events related in a single past episode, characterized by a unity of time or action and linked together by time markers, that is, events are reported as being temporally or causally consistent and depict a specific situation (Bangerter et al., 2014). Interrater agreement (two coders) of this coding scheme was previously shown to be high based on double-coding of 40 responses in a similar corpus (see Tescari et al., 2020), Cohen's kappa = 0.75 for story/not story. The transcripts featuring stories were further segmented into utterances and each utterance was coded for the STAR narrative elements by several different coders. We collapsed task and action into one category because they were difficult to distinguish. Interrater agreement was previously established (see Tescari et al., 2020) as high, based on double-coding by two different coders of another, similar data set. Because narrative elements are counts, interrater agreement is expressed as correlation coefficients (all *r*s for S, T/A, and *R* > 0.77).

5.3.3 | Features: Automatic content analysis using LIWC

Automatic content analysis of participants' answers was performed using LIWC 2015 (Pennebaker et al., 2015). We used the French LIWC dictionary (Piolat et al., 2011), which provides coding of words in 64 categories, including, for example, emotions, social words, cognitive mechanisms, or personal concerns. Each word can be coded in several categories. LIWC returns, for each applicant's response, the proportion of words in each of the categories. LIWC also returns other categories, such as the number of words the number of words of six letters or more, or the percentage of words featured in the used dictionary.

5.3.4 | Features: TF-IDF

TF-IDF is a statistic that reflects the importance of a term in the document (here, a document is the transcript of each participant's response). TF-IDF weights the frequency of the different terms in each document by their inverse frequency in the corpus (Salton & Buckley, 1988). The term-document matrix, which describes the frequency of terms (columns) in each document (rows), was computed from the texts constituted by participants' responses. Terms were words, compound words, and expressions typically written without space (e.g., *est-ce* in French). We trimmed the terms to be used: Terms with a frequency higher than 90% or lower than 5% of the number of documents were removed. Thus, the probability of a word being included in the trimmed term-document matrix was independent of the frequency of other words. Finally, TF-IDF measures were computed for each remaining term (Salton & Buckley, 1988). Note that TF-IDF features can potentially overlap with the LIWC features described in the previous subsection. For example, "ce" (*this*) is both a TF-IDF feature and included in the LIWC category of PRONOUN. However, the actual overlap between these sets of features is low. The average of the absolute values of all correlations between all TF-IDF features and all LIWC categories is *r* = 0.047.

5.4 | Analyses

Our primary analyses consist of using machine learning algorithms to predict the labels from the features (see Figure 1). We are thus in the domain of supervised machine learning (Liem et al., 2018). In supervised machine learning, predicting binary variables constitutes a classification task, whereas the prediction of a numerical variable constitutes a regression task (Mayor, 2015). We used 10-fold nested cross-validation for estimating the performance of models. Nested cross-validation separates information used for hyperparameter tuning from information about model accuracy, hence reducing the risk of overestimating accuracy (Hickman, Bosch, et al., 2022). Because the goal of machine learning analyses is to maximize predictive quality (Liem et al., 2018), we tested different algorithms and report results for those that performed best.

5.4.1 | Prediction of storytelling

This is a classification task (Liem et al., 2018): Predicting the binary classification of whether a response contained a story or not from the features. We combined LIWC and TF-IDF features with word count of the response and number of words >6 letters as additional features and trained machine learning models on them. We used support vector machines and random forest algorithms in Python using the Scikit-Learn package (Pedregosa et al., 2011). Standard machine learning practices (including preprocessing of data with z-score normalization) and nested 10-fold cross-validation as described above were applied. To understand the contribution of various features towards a classification result, we used the variable importance (varImp) metric provided by Scikit-learn.

5.4.2 | Prediction of STAR counts

This is a regression task (Liem et al., 2018): Predicting the counts of situation (S), task/action (T/A), or results (R) from features. We report results for three different combinations of features: (1) LIWC features, (2) TF-IDF features, and (3) both (all features). We ran analyses in R using the procedure for 10-fold nested cross-validation provided in Appendix A of Hickman, Bosch, et al. (2022) after removing cases with missing values. We report results from the three algorithms which fared best in analyses: random forests (caret

method: parRF), bagged CART; (caret method: treebag), and partial least squares (Abdi, 2010; caret method).

6 | RESULTS

6.1 | Prediction of storytelling

Manual coding revealed that participants produced stories in response to past-behavior questions 63% of the time. In Table 1, we report five performance measures for both algorithms for each fold and overall (mean). The first is accuracy, which is the percentage of correct classification of a response. Accuracy may be less informative when the data is skewed, for example, when a binary distribution differs strongly from parity. In such cases, a naïve classifier which simply classifies all cases as positives will automatically attain high accuracy (e.g., in our data, such a classifier would correctly identify stories 63% of the time). Thus, other performance measures are potentially relevant. The second performance measure is precision or positive predictive value, that is, the number of true positives (i.e., responses classified as stories by the human coders) divided by the number of positive classifications by the algorithm (which is the number of true positives plus the number of false positives). Precision is useful when the cost of false positive identifications (classifying a response as a story when it is not) is high. The third performance measure is recall or the true positive rate, which is the number of true positives divided by the sum of true positives and false negatives. Recall is useful when the cost of false negative identifications (classifying a response as not a story when it is) is high. Optimizing both precision and recall is the purpose of the fourth performance measure, F1, which is the harmonic mean of precision and recall (Sokolova & Lapalme, 2009). In our case, there is no reason to assume the costs of false negatives and false positives are different. Finally, we also report specificity, which is the proportion of negative predictions that are correct. This metric is relevant in training applications where corrective feedback would be given to applicants who do not produce a story.

In our data, random forest overall accuracy was 78%, and support vector machine overall accuracy was 70%. Random forest accuracy is thus substantially higher than the naïve classifier accuracy of 63%, while support vector machine accuracy is slightly higher. Overall, random forest performs better (but support vector machine is better on specificity). Specificity has the lowest values.

Table 2 depicts the features that contributed most towards the classification of story/not story. Specific French stems are indicated in italics, with their English translations (if the stems are translatable) in parentheses. LIWC categories are indicated in capitals and correspond to the English names for the French categories used in the analyses (e.g., THIRD PERS SINGULAR is a LIWC category abbreviation comprising third-person singular terms like *she*, *her*, or *him*). For both algorithms, past-tense verb conjugations like *ai* (had), *dû* (had to), or *était* (were) were among the features most predictive of storytelling, followed by numbers. Further predictive features include auxiliary verbs and pronouns (*je*, as evidenced in the stem *j*), which are related to a "dynamic" language style featuring personal narratives (Pennebaker et al., 2014) and social, cognitive, and affective processes.

6.2 | Prediction of STAR counts

Manual coding revealed that stories contained on average 8.1 utterances ($SD = 5.9$) describing the situation, 6.8 utterances ($SD = 5.8$) describing tasks/action, and 2.2 utterances ($SD = 2.4$) describing results. Table 3 displays performance measures for each algorithm for predicting the

TABLE 1 Performance measures for predicting storytelling by fold and overall for random forest and support vector machines, using all features.

	Random forest					Support vector machine				
Fold	Acc	Prec	Recall	F1	Spec	Acc	Prec	Rec	F1	Spec
0	0.78	0.76	0.92	0.83	0.59	0.78	0.83	0.79	0.81	0.76
1	0.78	0.74	1.00	0.85	0.40	0.71	0.73	0.85	0.79	0.47
2	0.85	0.87	0.93	0.90	0.67	0.63	0.79	0.65	0.72	0.58
3	0.83	0.77	1.00	0.87	0.61	0.71	0.69	0.87	0.77	0.50
4	0.70	0.79	0.79	0.79	0.50	0.65	0.77	0.71	0.74	0.50
5	0.70	0.63	0.95	0.76	0.45	0.70	0.68	0.75	0.71	0.65
6	0.85	0.86	0.97	0.91	0.44	0.82	0.87	0.90	0.89	0.56
7	0.78	0.69	0.95	0.80	0.62	0.63	0.58	0.74	0.65	0.52
8	0.80	0.79	0.96	0.87	0.42	0.73	0.87	0.71	0.78	0.75
9	0.75	0.83	0.83	0.83	0.55	0.65	0.78	0.72	0.75	0.45
Mean	0.78	0.77	0.93	0.84	0.52	0.70	0.76	0.77	0.76	0.57

Abbreviations: Acc, accuracy; prec, precision; spec, specificity.

TABLE 2 Features and weights for predicting storytelling for random forest and support vector machines.

Random forest		Support vector machine	
Feature	**Weight**	**Feature**	**Weight**
PAST TENSE	0.0478	*dû* (had to)	1.5929
était (was)	0.0410	NUMBER	1.0988
et (and)	0.0183	*a* (has)	1.0124
avait (had)	0.0169	*peu*	0.9661
WORD COUNT	0.0150	*etud*	0.9501
j	0.0144	INSIGHT	0.9373
AFFECTIVE PROCESSES	0.0139	SOCIAL PROCESSES	0.8766
PRESENT TENSE	0.0132	*dire* (say)	0.8439
pour (for)	0.0132	*et* (and)	0.8415
a (has)	0.0126	*travaille* (work)	0.8030
NUMBER	0.0116	*passer* (pass)	−0.7458
dû (had to)	0.0112	*quotidien* (everyday)	−0.7529
ai (have)	0.0111	*puisqu* (since)	−0.7626
SPACE	0.0097	*les* (the)	−0.7922
CERTAINTY	0.0094	TIME	−0.8369
VERBS	0.0092	*differente* (different)	−0.9030
EXCLUSIVE	0.0091	*voir* (see)	−0.9765
TENTATIVE	0.0090	*où* (where)	−1.0438
MOVEMENT	0.0089	*dossier* (dossier)	−1.0449
TIME	0.0088	*foi*	−1.1434

Note: Original French stems in italics, English translations in parentheses wherever the stem is interpretable, LIWC categories (English) in capitals.

utterance count of situation, task/action, and result, for LIWC features, TF-IDF features, and all features combined. As performance measures, we report root mean square error (RMSE), Pearson's *r*, and mean absolute error (MAE). Performance is higher when *r* increases, and when RMSE and MAE are minimized. Table 3 suggests the best performance for predicting the number of situation or task/action utterances is attained for TF-IDF features and PLSs. The best performance for predicting the number of results utterances is attained for all features using bagged CART. Moreover, the substantial *r* values for predicting the number of situation and task/action utterances suggest that predictive performance for these utterance types is better than for utterances about results.

Tables 4–6 display feature importance for algorithms (using all features combined) for predicting the situation (Table 4), task/action (Table 5), and results (Table 6) utterance counts. Features predictive of situation utterances include *et*, *que*, articles, conjunctions (as a LIWC category as well as single exemplars like *donc*), but also words related to cognitive processes, the past tense, and anxiety. Features predictive of task/action utterances include *et*, *de*, articles, conjunctions (as a LIWC category as well as single exemplars like *donc*), as well as causal language, which is present as a LIWC category, but also evident in constituents of phrases like *pour que* (to). Features predictive of results utterances include temporal adverbs (e.g., *quand*), affective processes, expressions of success (*réussi*), as well as words related to relativity in LIWC (movement, time, and space).

7 | DISCUSSION

The goal of the current study was to contribute to the automatic analysis of applicant verbal behavior in structured behavioral interviews for training purposes. Because the core feature of behavioral

TABLE 3 Performance of different algorithms in the prediction of situation, task/action, and result (STAR), using LIWC features, TF-IDF features, and all features.

	Situation			Task/Action			Result		
	RMSE	***r***	**MAE**	**RMSE**	***r***	**MAE**	**RMSE**	***r***	**MAE**
LIWC									
Bagged CART	5.498	0.361	4.132	4.956	0.458	3.558	2.426	0.118	1.924
Random Forests	5.530	0.351	4.126	5.096	0.503	3.666	2.396	0.086	1.901
PLS	6.068	0.187	4.688	5.780	0.246	4.214	2.439	0.094	1.925
TF-IDF									
Bagged CART	4.865	0.550	3.630	4.416	0.610	3.131	2.165	0.399	1.646
Random Forests	4.901	0.586	3.574	4.370	0.647	3.052	2.244	0.325	1.723
PLS	4.637	**0.602**	3.447	4.052	**0.667**	2.839	2.236	0.368	1.683
All features									
Bagged CART	4.875	0.546	3.640	4.429	0.607	3.168	2.137	**0.421**	1.650
Random Forests	4.946	0.559	3.658	4.451	0.641	3.149	2.220	0.336	1.722
PLS	4.934	0.537	3.748	4.206	0.651	3.034	2.376	0.307	1.834

Note: Highest *r* displayed in bold.

Abbreviations: LIWC, linguistic inquiry and word count; MAE, mean absolute error; PLS, partial least squares; RMSE, root mean square error; TF-IDF, term frequency-inverse document frequency.

TABLE 4 Feature importance for predicting utterance counts about situation (all features models).

Bagged classification and regression trees		Random forest		Partial least squares	
et (and)	92.178	*aussi* (also)	79.386	ARTICLES	98.992
que (that)	85.818	*et* (and)	77.043	CONJUNCTIONS	76.576
de (of)	71.999	*travailler* (work)	76.864	COGNITIVE PROCESSES	73.313
était (was)	57.673	*quelqu* (some)	76.151	*ben* (well)	72.482
PAST TENSE	53.127	ANXIETY	76.011	RELATIVITY	69.901
c	46.997	*de* (of)	75.895	*ils* (they)	66.449
NEGATIVE EMOTION	43.528	*que* (that)	75.604	*était* (was)	54.903
à (to)	42.602	*avant* (before)	75.029	voilà (there)	53.988
on (one/we)	39.719	*ils* (they)	73.340	VERBS	52.787
THIRD PERS PLURAL	39.513	*mai*	71.857	PREPOSITIONS	52.548
FIRST PERS SINGULAR	38.933	*mettre* (put)	71.252	*effectif* (effectiv)	51.937
HEARING	38.446	*le* (the)	70.298	*vrai* (true)	51.348
PREPOSITIONS	38.251	*était* (was)	70.204	*gen*	50.339
AUXILIARY VERBS	37.525	*ou* (or)	69.880	*trè*	50.254
donc (so)	37.333	*voilà* (there)	69.833	*pui*	48.754
CAUSATION	36.569	*ça* (that)	69.092	NONFLUENCIES	48.584
THIRD PERS SINGULAR	35.638	*y* (to)	69.008	*elle* (she)	46.155
le (the)	35.600	*donc* (so)	68.684	*là* (there)	46.076
FUNCTION WORDS	34.974	*euh* (uh)	68.318	*ça* (that)	44.716
PERS PRONOUNS	34.361	il (he)	67.377	*quelqu* (some)	44.098

Note: Original French stems in italics, English translations in parentheses wherever the stem is interpretable, LIWC categories (English) in capitals.

interviewing is the type of questions asked of applicants, especially past-behavior questions, we focused on applicants' responses to these questions. We thus implemented a suite of algorithms to detect the presence of absence of stories and their narrative elements (descriptions of situation, task/action, and results) from features automatically coded in transcribed speech.

Our findings demonstrate the feasibility of machine learning for identifying storytelling, with random forest producing 78% accuracy. Among other performance measures, specificity, or the proportion of negative predictions, was rather low. Both the presence or absence of stories in responses to past-behavior questions and the prevalence of three types of narrative elements (STAR) are predictable from a relatively arbitrary set of initial features. We found that both random forest and support vector machine algorithms worked well for story prediction. Further, for predicting the counts of STAR elements, PLS using TF-IDF features worked best for utterances about the situation and about tasks/action, whereas bagged CART worked best for predicting utterances about results.

Our findings have implications for applicant coaching in behavioral interviews. Because applicants may not be used to situational questions or past-behavior questions, they may not be able to provide appropriate answers. Indeed, storytelling responses to past-behavior questions are difficult to produce on-the-fly (Brosy et al., 2016) and thus occur infrequently (Bangerter et al., 2014; Brosy et al., 2020), with participants often producing less appropriate responses like pseudostories or describing their traits, beliefs, or opinions. This in turn may reduce their hiring recommendations (Bangerter et al., 2014). Interview coaching programs help applicants tell more effective stories in response to past-behavior questions, for example, by training them to construct stories that fulfill effectiveness criteria like consistency, relevance, level of detail, and the like (Ralston et al., 2003) or by training them to employ the STAR rule in constructing stories (Lukacik et al., 2022; Tross & Maurer, 2008). Such programs can increase applicant performance (Maurer & Solamon, 2006) and interview validity (Maurer et al., 2008). But they are costly to implement, requiring access to a coach to deliver tailored feedback. Automatically identified parameters of applicant responses may serve as a basis for communicating such tailored feedback to applicants about their performance, thus partially replacing a coach, or complementing a coach's activity. For example, a system could provide simple verbal feedback advising applicants to "talk about a specific episode you experienced" in case their initial response is identified as not being a story, or to emphasize certain aspects of the story, e.g., "describe in more detail what the task was

TABLE 5 Feature importance for predicting for predicting utterance counts about task/action (all features models).

Bagged classification and regression trees		Random forest		Partial least squares	
de (of)	96.373	*les* (the)	84.128	COGNITIVE PROCESS	91.597
donc (so)	83.283	*de* (of)	84.038	*été* (been)	88.223
et (and)	72.544	*donc* (so)	81.069	*collaborer* (collaborate)	85.462
les (the)	65.356	*et* (and)	79.210	ARTICLES	82.033
on (one/we)	46.945	*qu*	75.835	*est.c* (is that)	79.667
MOVEMENT	44.426	*l*	74.057	VERBS	74.635
des (some)	41.386	*des* (some)	72.989	CONJUNCTIONS	67.377
qu	38.110	*que* (that)	70.119	FUNCTION WORDS	65.113
à (to)	37.651	*qui* (who)	69.757	*se* (him/herself)	64.408
qui (who)	35.284	*est* (is)	67.796	CAUSATION	64.401
CAUSATION	32.832	*est.c*	67.161	PERS PRONOUNS	64.186
l	32.406	*on* (one/we)	66.721	*group* (group)	63.043
ADVERB	30.781	*a* (has)	66.134	cas (case)	62.546
que (that)	30.503	MOVEMENT	66.054	*ça* (that)	62.276
RELATIVITY	29.333	*pour* (for)	65.982	PRONOUNS	62.130
en (in)	29.133	*euh* (uh)	65.282	TIME	61.658
WORK	28.246	*equip* (equip)	64.530	NONFLUENCIES	61.099
tout (all)	28.080	RELATIVITY	64.340	FIRST PERS SINGULAR	60.656
euh (uh)	27.651	*ce* (this)	63.841	*ils* (they)	59.486
ACHIEVEMENT	26.342	*etaient* (were)	63.346	*avoi*	59.326

Note: Original French stems in italics, English translations in parentheses wherever the stem is interpretable, LIWC categories (English) in capitals.

and what you actually did" if they produced many more situation utterances than task/action utterances, as is often the case (Bangerter et al., 2014).

Future research might explore the efficacy of such brief interventions on applicant storytelling behavior and interview performance, either in a traditional format or a AVI format. Testing effects of different interventions based on the model predictions is especially important because the predictive performance of the current models is not perfect. Considering the relatively low specificity values (0.52 for random forest and 0.58 for support vector machine), negative predictions about storytelling may often be incorrect. Applicants who respond with a story but whose responses are misclassifed and who are thus advised to "talk about a specific episode you experienced" may react with confusion. Further, predictions of STAR utterances would need to be translated into specific feedback to be effective, for example, by determining a threshold for the number of utterances to trigger advice to participants to produce more utterances of a specific kind.

Future research might also attempt to improve the predictive power of the current analyses. While random forest performs well for small sample sizes (Qi, 2012), training algorithms on larger data sets, or using more powerful (e.g., deep learning) algorithms or a combination of both could substantially improve the quality of our predictions.

Future research might also explore the utility of devising question-specific models. In our study, we did not have enough data to do this. However, specific questions create contexts for answers, and storytelling answers may be more easily detectable based on question-specific words rather than generic words. However, such analyses require large datasets and potentially limit the utility of the findings to a specific question, which may only be feasible for organizations with a high volume of applicants.

There are some limitations to this study. First, we based our analyses on transcripts. In a fully fledged prediction pipeline, it would be desirable to implement automatic speech recognition (ASR) to eliminate time-intensive manual transcription. It is unknown how the current predictive performance of storytelling and STAR narrative elements would change when based on automatically transcribed speech. Future research might use an ASR component to generate the features for the current models. Another limitation is the focus on exclusively textual features to predict storytelling. Because storytelling is a multimodal phenomenon (Bangerter et al., 2011; Bavelas et al., 2014; Okada et al., 2016), including prosodic or visual features in a predictive model might increase performance. Indeed, low-level

TABLE 6 Feature importance for predicting utterance counts about results (all features models).

Bagged classification and regression trees		Random forest		Partial least squares	
à (to)	90.689	*reussi* (succeeded)	83.834	FUNCTION	95.495
était (was)	83.788	*à* (to)	82.138	NONFLUENCIES	94.190
AFFECTIVE PROCESSES	72.778	*était* (was)	81.555	*elle* (she)	68.725
TIME	67.656	*quand* (when)	71.642	ADVERBS	67.522
AUXILIARY VERBS	66.223	*sort*	66.328	*pu* (could)	66.751
quand (when)	65.869	*cette* (this)	65.433	PRONOUNS	62.036
et (and)	65.475	*I*	64.720	PERS PRONOUNS	61.707
MOVEMENT	61.255	*premier* (first)	64.692	AFFECTIVE PROCESSES	61.556
on (one/we)	59.907	*et* (and)	64.283	*était* (was)	57.394
VERBS	57.037	*pour* (for)	64.056	COGNITIVE PROCESS	55.408
a (has)	55.883	THIRD PERS PLURAL	64.020	*été* (was)	53.998
ARTICLES	55.562	*ont* (had)	62.989	*quand* (when)	52.690
SPACE	55.489	*a* (had)	62.894	*se* (him/herself)	50.155
reussi (succeeded)	55.299	*rôle* (role)	62.743	PRESENT TENSE	48.574
PRESENT TENSE	53.234	*comment* (how)	62.683	*client* (client)	47.808
NEGATIVE EMOTION	52.799	*ans* (years)	62.627	TIME	47.692
au (to)	51.750	*part*	62.466	SPACE	47.671
de (of)	51.242	*ne* (not)	62.346	ARTICLE	47.443
PREPOSITIONS	51.185	*pas* (not)	62.066	*bien* (good)	46.087
foi	47.139	*plutôt* (rather)	61.248	RELATIVITY	43.483

Note: Original French stems in italics, English translations in parentheses wherever the stem is interpretable, LIWC categories (English) in capitals.

nonverbal and paraverbal features could potentially be useful as indicators of responses like storytelling. For example, Nguyen et al. (2014) found in their study (where applicants responded to past-behavior questions) that both longer speaking time and longer turn duration in applicants and interviewer backchannel behavior correlated with hireability ratings, but did not analyze whether participants told stories or not. In a separate analysis of the same data set, Bangerter et al. (2014) coded storytelling and found that it correlated with the same hireability ratings. It is thus plausible that the speech-related and visual features predictive of hireability ratings in Nguyen et al. (2014) represent nonverbal correlates of storytelling activity, which involves narrators (applicants) taking extended turns and audiences (recruiters) adopting listening roles (hence their production of backchannels). Combining these features with the textual features we investigated might increase predictive performance.

Despite these limitations, the current study contributes to developing machine learning approaches to analyze applicant verbal behavior within the framework of behavioral interviewing. By enabling automatic identification of a key dimension of responses to past-behavior questions, namely storytelling, we enable better content-based analyses of applicants' verbal responses. This in turn may contributes to designing feedback delivery systems for interview coaching that help them to better tell their stories, and thus ultimately improve the functioning of behavioral interviews.

ACKNOWLEDGMENTS

This study was funded by Swiss National Science Foundation grant 10DL2C_183065. Open access funding provided by Universite de Neuchatel.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Eric Mayor http://orcid.org/0000-0001-9441-7592

REFERENCES

Abdi, H. (2010). Partial least squares regression and projection on latent structure regression (PLS Regression). *WIREs Computational Statistics*, *2*(1), 97–106.

Bangerter, A., Corvalan, P., & Cavin, C. (2014). Storytelling in the selection interview? How applicants respond to past behavior questions. *Journal of Business and Psychology*, *29*(4), 593–604.

Bangerter, A., Mayor, E., & Doehler, S. P. (2011). Reported speech in conversational storytelling during nursing shift handover meetings. *Discourse Processes*, *48*(3), 183–214.

Bavelas, J., Gerwing, J., & Healing, S. (2014). Effect of dialogue on demonstrations: Direct quotations, facial portrayals, hand gestures, and figurative references. *Discourse Processes*, *51*(8), 619–655.

Benesty, M. (2019). *Unine: Unine light stemmer_. R package version 0.2.0*. https://CRAN.R-project.org/package=unine

Biel, J. I., Tsiminaki, V., Dines, J., & Gatica-Perez, D. (2013, December). *Hi YouTube! personality impressions and verbal content in social video*. Proceedings of the 15th ACM international conference on multimodal interaction, 119–126.

Brosy, J., Bangerter, A., & Mayor, E. (2016). Disfluent responses to job interview questions and what they entail. *Discourse Processes*, *53*(5–6), 371–391.

Brosy, J., Bangerter, A., & Ribeiro, S. (2020). Encouraging the production of narrative responses to past-behaviour interview questions: Effects of probing and information. *European Journal of Work and Organizational Psychology*, *29*(3), 330–343.

Burnett, J. R., & Motowidlo, S. J. (1998). Relations between different sources of information in the structured selection interview. *Personnel Psychology*, *51*(4), 963–983.

Chen, L., Zhao, R., Leong, C. W., Lehman, B., Feng, G., & Hoque, M. E. (2017, October). Automated video interview judgment on a large-sized corpus collected online. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)* (pp. 504–509). IEEE.

Gebhard, P., Schneeberger, T., Andre, E., Baur, T., Damian, I., Mehlmann, G., Konig, C., & Langer, M. (2019). Serious games for training social skills in job interviews. *IEEE Transactions on Games*, *11*(4), 340–351.

Heimerl, A., Mertes, S., Schneeberger, T., Baur, T., Liu, A., Becker, L., & André, E. (2022). "GAN I hire you?"—A system for personalized virtual job interview training. arXiv preprint arXiv, 2206.03869.

Hickman, L., Bosch, N., Ng, V., Saef, R., Tay, L., & Woo, S. E. (2022). Automated video interview personality assessments: Reliability, validity, and generalizability investigations. *Journal of Applied Psychology*, *107*(8), 1323–1351.

Hickman, L., Thapa, S., Tay, L., Cao, M., & Srinivasan, P. (2022). Text preprocessing for text mining in organizational research: Review and recommendations. *Organizational Research Methods*, *25*(1), 114–146.

Hollandsworth, J. G., Kazelskis, R., Stevens, J., & Dressel, M. E. (1979). Relative contributions of verbal, articulative, and nonverbal communication to employment decisions in the job interview setting. *Personnel Psychology*, *32*(2), 359–367.

Hoque, M., Courgeon, M., Martin, J. C., Mutlu, B., & Picard, R. W. (2013, September). Mach: My automated conversation coach. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (pp. 697–706).

Huffcutt, A. I., & Arthur, W., Jr. (1994). Hunter and Hunter (1984) revisited: Interview validity for entry level jobs. *Journal of Applied Psychology*, *79*, 184–190.

Huffcutt, A. I., Conway, J. M., Roth, P. L., & Stone, N. J. (2001). Identification and meta-analytic assessment of psychological constructs measured in employment interviews. *Journal of Applied Psychology*, *86*(5), 897–913.

Huffcutt, A. I., Culbertson, S. S., Goebl, A. P., & Toidze, I. (2017). The influence of cognitive ability on interviewee performance in traditional versus relaxed behavior description interview formats. *European Management Journal*, *35*(3), 383–387.

Kantrowitz, T. M., Tuzinski, K. A., & Raines, J. M. (2018). *2018 Global assessment trends report*. SHL.

Kessler, R. (2006). *Competency-based interviews*. Career Press.

Klehe, U. C., & Latham, G. (2006). What would you do—Really or ideally? Constructs underlying the behavior description interview and the situational interview in predicting typical versus maximum performance. *Human Performance*, *19*(4), 357–382.

Langer, M., König, C. J., Gebhard, P., & André, E. (2016). Dear computer, teach me manners: Testing virtual employment interview training. *International Journal of Selection and Assessment*, *24*(4), 312–323.

Levashina, J., Hartwell, C. J., Morgeson, F. P., & Campion, M. A. (2014). The structured employment interview: Narrative and quantitative review of the research literature. *Personnel Psychology*, *67*(1), 241–293.

Liem, C. C., Langer, M., Demetriou, A., Hiemstra, A. M., Wicaksana, A. S., Born, M. P., & König, C. J. (2018). Psychology meets machine learning: Interdisciplinary perspectives on algorithmic job candidate screening. In H. Escalante (Ed.), *Explainable and interpretable models in computer vision and machine learning. The Springer series on challenges in machine learning* (pp. 197–253). Springer.

Lin-Stephens, S., Manuguerra, M., Tsai, P. J., & Athanasou, J. A. (2022). Stories of employability: Improving interview narratives with image-supported past-behaviour storytelling training. *Education+ Training*, *64*, 577–597.

Lukacik, E. R., Bourdage, J. S., & Roulin, N. (2022). Into the void: A conceptual model and research agenda for the design and use of asynchronous video interviews. *Human Resource Management Review*, *32*(1), 100789.

Maurer, T. J., & Solamon, J. M. (2006). The science and practice of a structured employment interview coaching program. *Personnel Psychology*, *59*(2), 433–456.

Maurer, T. J., Solamon, J. M., & Lippstreu, M. (2008). How does coaching interviewees affect the validity of a structured interview? *Journal of Organizational Behavior*, *29*(3), 355–371.

Mayor, E. (2015). *Learning predictive analytics with R*. Packt Publishing Ltd.

Motowidlo, S. J. (1999). Asking about past behavior versus hypothetical behavior. In R. W. Eder & M. M. Harris (Eds.), *The employment interview handbook* (pp. 179–190). Sage.

Motowidlo, S. J., Carter, G. W., Dunnette, M. D., Tippins, N., Werner, S., Burnett, J. R., & Vaughan, M. J. (1992). Studies of the structured behavioral interview. *Journal of Applied Psychology*, *77*(5), 571–587.

Muralidhar, S., Nguyen, L. S., Frauendorfer, D., Odobez, J.-M., Schmid Mast, M., & Gatica-Perez, D. (2016, November). *Training on the job: Behavioral analysis of job interviews in hospitality*. Proceedings of ACM international conference on multimodal interaction (ICMI), Tokyo.

Naim, I., Tanveer, M. I., Gildea, D., & Hoque, M. E. (2018). Automated analysis and prediction of job interview performance. *IEEE Transactions on Affective Computing*, *9*(2), 191–204.

Nguyen, L. S., Frauendorfer, D., Mast, M. S., & Gatica-Perez, D. (2014). Hire me: Computational inference of hirability in employment interviews based on nonverbal behavior. *IEEE Transactions on Multimedia*, *16*(4), 1018–1031.

Okada, S., Hang, M., & Nitta, K. (2016). Predicting performance of collaborative storytelling using multimodal analysis. *IEICE TRANSACTIONS on Information and Systems*, *99*(6), 1462–1473.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2*. The University of Texas at Austin.

Pennebaker, J. W., Chung, C. K., Frazee, J., Lavergne, G. M., & Beaver, D. I. (2014). When small words foretell academic success: The case of college admissions essays. *PLoS One*, *9*(12), e115844.

Piolat, A., Booth, R. J., Chung, C. K., Davids, M., & Pennebaker, J. W. (2011). La version française du dictionnaire pour le LIWC: Modalités de construction et exemples d'utilisation. *Psychologie Française*, *56*(3), 145–159.

Qi, Y. (2012). Random forest for bioinformatics. In C. Zhang & Y. Ma (Eds.), *Ensemble machine learning: Methods and applications* (pp. 307–323). Springer US.

Ralston, S. M., Kirkwood, W. G., & Burant, P. A. (2003). Helping interviewees tell their stories. *Business Communication Quarterly*, *66*(3), 8–22.

Ramos, J. (2003, December). Using TF-IDF to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning* (pp. 29–48).

Rasmussen, K. G. (1984). Nonverbal behavior, verbal behavior, resumé credentials, and selection interview outcomes. *Journal of Applied Psychology*, *69*(4), 551–556. https://doi.org/10.1037/0021-9010.69.4.551

Roulin, N. (2017). *The psychology of job interviews*. Routledge.

Roulin, N., Bangerter, A., & Wüthrich, U. (2012). *Réussir l'entretien d'embauche comportemental: La méthode pour identifier et sélectionner les futurs employés performants*. De Boeck Professionals.

Rupasinghe, A. T., Gunawardena, N. L., Shujan, S., & Atukorale, D. A. S. (2016, September). Scaling personality traits of interviewees in an online job interview by vocal spectrum and facial cue analysis. In *2016 Sixteenth International Conference on Advances in ICT for Emerging Regions (ICTer)* (pp. 288–295). IEEE

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, *24*, 513–523.

Savoy, J. (1999). A stemming procedure and stopword list for general French corpora. *Journal of the American Society for Information Science*, *50*(10), 944–952. https://doi.org/10.1002/(SICI)1097-4571(1999)50:10%3C944::AID-ASI9%3E3.0.CO;2-Q

Sharma, A., & Grant, D. (2011). Narrative, drama and charismatic leadership: The case of Apple's Steve jobs. *Leadership*, *7*(1), 3–26.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, *45*(4), 427–437.

Stevens, C. K., & Kristof, A. L. (1995). Making the right impression: A field study of applicant impression management during job interviews. *Journal of Applied Psychology*, *80*(5), 587–606.

Suen, H. Y., Hung, K. E., & Lin, C. L. (2020). Intelligent video interview agent used to predict communication skill and perceived personality traits. *Human-Centric Computing and Information Sciences*, *10*(1), 3.

Taylor, P. J., & Small, B. (2002). Asking applicants what they would do versus what they did do: A meta-analytic comparison of situational and past behaviour employment interview questions. *Journal of Occupational and Organizational Psychology*, *75*(3), 277–294.

Tescari, M. E., Bangerter, A., Györkös, C., Padoan, C., Fasel, S., Nicolier, L., Hondius, L., & Hitz, L. (2020). *Storytelling in the job interview: How do information and professional experience influence production of stories and interviewee performance? Unpublished manuscript*. University of Neuchâtel.

Tross, S. A., & Maurer, T. J. (2008). The effect of coaching interviewees on subsequent interview performance in structured experience-based interviews. *Journal of Occupational and Organizational Psychology*, *81*(4), 589–605.

Turner, T. S. (2004). *Behavioral interview guide: A practical, structured approach for conducting effective selection interviews*. Trafford Publishing.

Wiesner, W. H., & Cronshaw, S. F. (1988). A meta-analytic investigation of the impact of interview format and degree of structure on the validity of the employment interview. *Journal of Occupational Psychology*, *61*(4), 275–290.

How to cite this article: Bangerter, A., Mayor, E., Muralidhar, S., Kleinlogel, E. P., Gatica-Perez, D., & Schmid Mast, M. (2023). Automatic identification of storytelling responses to past-behavior interview questions via machine learning. *International Journal of Selection and Assessment*, *31*, 376–387. https://doi.org/10.1111/ijsa.12428