Entretien vidéo différé : modèle prédictif pour la pré-sélection de candidats sur la base du contenu verbal

Léo Hemamou[123], Grégory Wajntrob[1], Jean-Claude Martin[2], Chloé Clavel[3]

[1]EASYRECRUE, 3 bis rue de la Chaussée d'Antin, 75009 Paris

[2]LIMSI, CNRS, Université Paris-Saclay, Bât 508, rue John von Neumann, Campus Universitaire, F-91405 Orsay

[3]LTCI-CNRS, Telecom-Paristech, Université Paris-Saclay, 75013 Paris

RÉSUMÉ

Les entretiens vidéo différés sont devenus plus en plus populaires dans le milieu des ressources humaines et constituent un objet de recherche en traitement informatique de signaux sociaux. Dans cet article, nous passons en revue plusieurs études sur les recherches pertinentes en termes d'analyses comportementales et de systèmes d'entraînement à passer des entretiens d'embauche. Nous présentons ensuite un premier modèle de prédiction automatiques des classements des candidats pour un recruteur. Nous nous focalisons dans cette première étude sur l'analyse automatique du contenu verbal d'entretiens vidéo différés. Ce travail s'inscrit en partenariat avec la société EASYRECRUE, société proposant une plateforme d'entretiens vidéo différés, avec laquelle un corpus de plus de 300 candidats évalués par un recruteur a été collecté. Nos travaux ont montré la faisabilité de l'exploitation du contenu verbal après l'utilisation d'un outil de reconnaissance automatique. Les perspectives en ce qui concerne le traitement automatique de la prosodie et des comportements non verbaux sont présentées.

Mots-Clés

Entretiens vidéo différés; Évaluation automatique; Informatique sociale; Psychologie sociale.

1. INTRODUCTION

Le développement des nouvelles technologies impacte tous les secteurs d'activités, y compris celui des Ressources Humaines. Ainsi, l'entretien vidéo différé permet d'organiser en asynchrone des entretiens avec des candidats et de les évaluer. Les candidats se connectent à une plateforme, se filment pendant qu'ils répondent à des questions définies par les recruteurs. La plateforme permet ensuite à plusieurs recruteurs d'évaluer le candidat, d'échanger entre eux et d'inviter éventuellement le candidat à un entretien en face à face. Les recruteurs établissent au préalable un questionnaire de recrutement et y associent des critères d'évaluation. Les questions posées durant l'entretien différé sont en général des questions sur la motivation, les expériences professionnelles, ou des questions de mises en situation. Le/la candidat-e reçoit une invitation pour répondre à ces questions et enregistre en vidéo ses réponses selon ses disponibilités dans un temps limité. Il ou elle n'a pas connaissance des questions à l'avance et ne peut pas se réenregistrer afin de préserver la spontanéité de ses réponses. Le recruteur reçoit les vidéos sur une interface et peut ainsi comparer et évaluer les différents profils avec ses équipes selon les critères définis précédemment. De plus en plus d'entreprises font le choix de ce type d'entretien comme outil de présélection. Ce choix est motivé par l'accès à un plus grand nombre et une plus grande diversité de candidats et la réduction du temps de traitement et de prise de rendez-vous. Le nombre de telles candidatures vidéo devient cependant de plus en plus volumineux et difficile à traiter « manuellement » par un ou deux recruteurs. Il devient donc nécessaire d'envisager une aide pour le recruteur devant traiter parfois plusieurs dizaines (voire centaines) d'entretiens vidéo. Aucune recherche, à notre connaissance, portant sur les entretiens vidéo différés n'a été effectuée dans un contexte hors laboratoire. Il peut y avoir pourtant de grandes différences entre une situation réelle d'embauche dans laquelle des candidats sont réellement motivés par une véritable candidature, et des conditions expérimentales contrôlées dans lesquelles les participants simulent un intérêt pour un poste fictif.

1.1 L'entretien d'embauche

Plusieurs méthodes pour évaluer l'adéquation d'un candidat à un emploi ou à l'entreprise existent. Ces outils peuvent être des tests classiques (personnalité, connaissances liées au poste), des vérifications de références, des évaluations par des pairs ou des entretiens [51]. Parmi ces outils, l'entretien d'embauche reste le moyen le plus utilisé afin d'évaluer des candidats. Un entretien permet à un recruteur de vérifier des informations, d'évaluer les compétences du candidat, de déterminer une personnalité ou de vérifier l'adéquation du candidat avec la culture de l'entreprise ou le poste. L'entretien peut s'effectuer via plusieurs canaux tels que le téléphone, la vidéo, le face à face, ou plus récemment l'entretien vidéo différé. Ces canaux, la structure et la construction de l'entretien ont une influence sur la prestation du candidat et sur l'évaluation qu'effectue le recruteur [53]. Ainsi, plus un entretien est structuré, plus sa validité prédictive augmente et la variabilité inter-recruteur diminue [11, 51]. La structure de l'entretien d'embauche améliore aussi l'équité des candidats [20, 27, 51]. Dans cet article, nous nous focalisons sur les entretiens vidéo différés. Par nature, les entretiens vidéo différés sont structurés. En effet, certaines caractéristiques de la structuration sont inhérentes à la conception de l'entretien différé comme l'usage systématique des mêmes questions pour tous les candidats à un poste, l'absence de discussion après chaque question, l'interdiction pour le candidat de poser des questions, et la limitation de l'interaction sociale durant l'entretien. D'autres caractéristiques dépendent de la construction même de l'entretien telles que l'élaboration de questions pertinentes et liées au poste, la construction d'un nombre suffisant de questions ou le contrôle de connaissances préalables tel que le curriculum vitae (CV). Les dernières caractéristiques dépendent directement de la méthode employée par les différents recruteurs affectés au poste pour la notation. La notation peut être faite pour chaque réponse à une question, ou au niveau global de l'entretien, elle peut être faite individuellement par chaque

recruteur ou collectivement par tous les recruteurs en charge du poste visé. Les recruteurs construisent l'évaluation d'un candidat par rapport à ses réponses verbales, mais aussi par rapport à leurs comportements non verbaux [17]. Les premières impressions influencent ainsi souvent l'évaluation d'un recruteur [36, 52]. Les candidats quant à eux peuvent user de techniques afin d'influencer positivement l'évaluation d'un recruteur grâce à des stratégies de gestion de l'impression ou de persuasion [39].

1.2 Le traitement automatique de signaux sociaux dans le contexte du recrutement

Le traitement informatique de signaux sociaux (social signal processing) fournit des méthodes et des outils servant à analyser automatiquement des signaux collectés dans de nombreuses situations. Nous pouvons citer des situations telles que l'évaluation de l'engagement d'étudiants à apprendre à distance, la détection de stress par webcam, ou la détection d'émotions [5, 16, 24]. Ces techniques ont aussi été utilisées dans le contexte du recrutement avec deux objectifs principaux : 1) l'aide aux candidats en les entraînant à passer des entretiens d'embauche ou à prendre la parole en public, et 2) l'aide au recruteur pour l'évaluation automatique de candidats. Des agents virtuels comme MACH [21] et TARDIS [2] ont été proposés afin d'aider des candidats à s'entraîner à passer des entretiens d'embauche. Des agents virtuels ont été construits notamment pour l'entraînement à la prise de parole en public [3] ou pour améliorer les capacités d'interactions [1, 54]. En complémentarité de la construction de ces agents virtuels, des outils de feedback automatique ont vu le jour tel que Automanner [55], un outil qui extrait automatiquement et avertit l'utilisateur de l'utilisation de gestes parasites, Rhema [56], un outil aidant des individus durant une présentation à parler à la bonne vitesse et à la bonne intensité, ou ROC Speak [59], une plateforme semi-automatisée donnant des retours d'informations lors d'une présentation vidéo grâce à une détection automatique des sourires ou du ton de la voix par exemple. Les travaux existants se limitent généralement à l'analyse d'entretiens d'embauche dont le poste à pourvoir est fictif. De plus, le contenu verbal est la modalité la moins étudiée et plusieurs travaux n'adoptent pas le même point de vue quant à l'apport de cette modalité [7, 31]. Dans cet article, nous présentons dans une première partie une étude de l'existant des travaux du traitement automatique des signaux sociaux dans le cadre des entretiens d'embauche et de la prise en parole en public. Nous exposerons ensuite dans une seconde partie notre apport à la communauté par la collecte d'entretiens vidéo différés puis dans une troisième et quatrième partie nous montrerons notre démarche d'extraction des descripteurs et la faisabilité d'un tri automatique de candidats basé sur le contenu verbal de leur réponse. Enfin, les deux dernières parties seront dédiées à l'étude des résultats et à la discussion des différents axes de recherche ouverts par cette étude.

2. ÉTUDE DE L'EXISTANT

Au cours de cette partie, nous présentons l'étude de l'existant du traitement automatique des signaux sociaux pour les entretiens d'embauche et la prise de parole en public. Nous décrivons les principaux comportements socio-émotionnels étudiés, les corpus construits, les indices multimodaux extraits et les méthodes d'analyse mises en place.

2.1 Comportements socio-émotionnels étudiés

Après une étude théorique, les auteurs de TARDIS [12] proposent quatre catégories d'informations échangées lors d'un entretien. L'attraction sociale est caractérisée par le montant d'appréciation qu'une personne peut provoquer chez les autres. L'engagement est le processus par lequel deux ou plusieurs participants établissent, maintiennent et terminent une relation perçue. La perception d'efficacité personnelle s'exprime souvent par une confiance dans la maîtrise des situations difficiles. L'état d'esprit réfère à une expression de la favorabilité ou non envers une personne ou un sujet particulier. Les expressions de cet état d'esprit considérées sont le stress, l'embarras, la sensation d'être mal à l'aise, l'ennui, la concentration, l'hésitation et le soulagement. Vis-à-vis des caractéristiques de l'évaluation des candidats, Nguyen et al. [35] proposent à la suite d'une analyse par composantes principales la décomposition des critères d'évaluations en trois catégories : les compétences sociales, les compétences de communication et les compétences professionnelles. Une grande partie des autres travaux s'articule autour de l'évaluation automatique de la performance globale, que ce soit dans un contexte d'entretien face à face [30, 32, 34] ou d'entretien vidéo différé [6, 7]. L'évaluation de la performance globale est parfois accompagnée d'autres critères tels que la persuasion [6, 34], la résistance au stress [6, 15, 32], le *leadership* [7, 30, 32] ,l'enthousiasme ou l'engagement. Plusieurs travaux s'intéressent à une composante spécifique telle que la communication orale [43–45] ou le stress [15]. Il a été aussi observé que la personnalité peut impacter l'évaluation du recruteur [51]. Plusieurs chercheurs orientent donc leurs travaux vers la détection de personnalité lors de monologues [4, 13, 14, 18] ou d'entretiens face à face [48]. Le mimétisme candidat-recruteur dans les entretiens d'embauche est étudié [26]. Les émotions dans les entretiens d'embauche ont été peu utilisées. Parmi celles-ci, seule la peur a montré des résultats intéressants [9]. Aussi, à notre connaissance, les effets de la valence, de l'activation physiologique ou de la dominance n'ont pas été étudiés. Dans cet article, seule la performance globale est évaluée, mais nous envisageons d'explorer prochainement les critères les plus utilisés parmi les recruteurs.

2.2 Corpus

A notre connaissance, un seul corpus d'entretien d'embauche réel (dont l'objectif est une mission à pourvoir) a été collecté et fait l'objet d'analyses automatiques. Ce corpus est constitué de 62 entretiens d'embauche face à face pour une mission marketing dont les candidats sont principalement des étudiants. Par ailleurs, nous pouvons séparer les autres corpus utilisés dans ce type de recherche en deux catégories qui sont les entretiens d'embauche non différés (face à face) et les entretiens d'embauches différés. Parmi les corpus d'entretiens d'embauches d'entraînement face à face nous pouvons citer deux corpus construits au Massachusetts Institute of Technology [21, 32] comprenant respectivement 138 et 28 étudiants, un corpus de 169 étudiants dans les services liés à l'accueil [30] et un corpus de 15 jeunes en insertion professionnelle [2]. De nombreux corpus d'entretiens vidéo différés ont aussi été constitués : un corpus de 36 employés [6], un corpus de 106 étudiants de l'université de Bangalore [43] et un corpus comprenant plus de 250 vidéos d'internautes récoltés en utilisant des outils de *crowdsourcing* [7]. Le corpus de ChaLearn composé de 10 000 vlogs d'une quinzaine de secondes dont le but initial était la prédiction de la personnalité lors des premières impressions a été complété par des annotations du type « la personne devrait-elle être

invitée à passer un entretien d'embauche ? ». Certains chercheurs s'intéressent aussi aux CV vidéo en ligne notamment Nguyen et al qui ont constitué un corpus de CV vidéo provenant de YouTube [35]. Cette collecte de grande base de données de vlogs et de CV vidéo a deux objectifs ; le premier objectif est de pouvoir tester la puissance prédictive d'outils automatique sur de grands corpus ;le deuxième objectif est d'obtenir une plus grande diversité au sein des candidats analysés. Enfin, des corpus de prise de parole en public pour des présentations professionnelles ou des mises en situations réelles ont aussi été constitués autour de l'entraînement au travail d'hôte d'accueil de 169 étudiants en école d'hôtellerie [29] ou l'entraînement aux présentations orales en milieu professionnel avec la participation de 36 employés [9].Certains corpus sont annotés par des experts ou des étudiants en psychologie [6, 7, 30, 34, 48]. D'autres corpus utilisent des outils de crowdsourcing [13, 32]. Enfin certains chercheurs utilisent uniquement des observateurs naïfs [45]. La table 1 regroupe un récapitulatif des corpus utilisés dans les précédents travaux. Cet article propose la constitution d'un corpus d'entretiens d'embauche différés de 305 candidats dont le poste à pourvoir est réel. Plus d'informations sont disponibles dans la partie 3 quant à la constitution de ce corpus.

Table 1. Tableau récapitulatif des corpus utilisés dans les précédents travaux

Travaux	Type du corpus	Poste réel à pourvoir	Nombre de candidats
[34]	Entretien face à face	Mission marketing	36
[30]	Entretien face à face	Non	169
[32]	Entretien vidéo différé	Non	138
[6]	Entretien face à face / Entretien vidéo différé	Non	36
[44]	Entretien face à face / Entretien vidéo	Non	106
[48]	Entretien vidéo différé	Non	36/8
[7]	Entretien vidéo différé	Non	260
[14]	Vlog	Non	3000
Cette étude	Entretien vidéo différé	Poste commercial	305

2.3 Descripteurs multimodaux

Nous pouvons séparer les indices étudiés en trois catégories qui sont 1) les indices visuels et les comportements non verbaux, 2) la prosodie, et 3) le contenu verbal. Parmi les comportements non verbaux, les indices les plus utilisés sont la direction du regard, l'utilisation des sourires et l'orientation de la tête. La proximité avec la caméra [2, 35], l'estimation de « l'énergie dégagée » par un candidat [34] ou sa posture [12] pendant son entretien sont aussi des indices utilisés. Des descripteurs de plus bas niveau sont aussi extraits tels que les unités d'action pour les expressions faciales [7, 45] ou des repères faciaux [32]. D'après la littérature en psychologie, l'apparence physique influence l'évaluation des recruteurs et plusieurs chercheurs ont donc annoté cette caractéristique [34, 45]. Concernant la prosodie, les descripteurs tels que la fréquence fondamentale, les temps de pause, l'intensité de la voix ou la vitesse de parole sont souvent utilisés [6, 30, 32]. De même que pour les comportements non verbaux, de nombreux descripteurs bas niveaux sont utilisés tels que les *Mel-frequency cepstral coefficients* ou des descripteurs du signal spectral et du signal sonore. A propos du contenu verbal, les dictionnaires (i.e. Linguistic Inquiry Word Count (LIWC) ou listes de mots parasites) et des statistiques lexicales (nombre de vocables divisé par le nombre de mots total communément appelé *Type Token Ratio* ou TTR, nombre de mots, nombres de mots de plus de 6 lettres, …) sont les approches les plus courantes [31, 32, 45]. Néanmoins, d'autres méthodes ont été utilisées afin d'extraire des descripteurs de texte comme la modélisation de thème par allocation de Dirichlet latente [32], la modélisation par sac de mots [7] ou la modélisation par *word embedding* avec l'algorithme Doc2Vec [6]. La table 2 propose une synthèse des différentes modalités utilisées dans les différents travaux considérés. Dans cet article, nous étudierons uniquement le contenu verbal. Il nous parait intéressant d'extraire des descripteurs permettant d'évaluer le registre de langue, la singularité lexicale, l'émotion véhiculée à travers les mots employés ou la sémantique utilisée en complémentarité des méthodes classiques de dictionnaires et de statistiques.

2.4 Méthodes d'analyse

Des analyses statistiques peuvent être effectuées entre les descripteurs extraits et les critères annotés [35, 45]. Les coefficients de corrélation intra-classes entre recruteurs renseignent sur la difficulté d'un critère à être perçu [35, 45]. Les modèles d'apprentissage supervisés sont les méthodes d'analyse les plus utilisées. Afin d'obtenir des descripteurs de taille fixe pour la construction de tels modèles des fonctions statistiques (moyenne, écart-type, premier quartile, troisième quartile) sont appliquées pour certains descripteurs. D'autres méthodes s'apparentant à du sequence mining [6] sont explorées. Parmi les méthodes d'apprentissage utilisées, nous pouvons citer les forêts aléatoires, les régressions logistiques, les séparateurs à vaste marge ou les réseaux bayésiens. L'analyse des modèles appris (poids attribué aux descripteurs, descripteurs sélectionnés) permet de mieux comprendre les critères utilisés par les recruteurs. Certains chercheurs utilisent uniquement des descripteurs de bas niveau pour entraîner des modèles d'apprentissage profond. Plusieurs architectures ont été construites, d'une part, pour répondre à l'enjeu de la temporalité, d'autre part pour répondre au problème de la multimodalité. Afin de répondre au problème de temporalité, les réseaux de neurones récurrents sont employés [13]. L'utilisation de réseaux à convolution pré-entraînés est envisagée afin d'extraire des représentations de descripteurs plus complexes [38]. Des chercheurs traitent les informations hors de la zone corporelle comme les pixels de l'environnement extérieur (chambre ou salon par exemple) pour essayer d'améliorer la prédiction de la personnalité [38]. Enfin, nous pouvons citer les travaux de Automanner [55] qui cherchent à détecter automatiquement et d'une façon non supervisée les gestes parasites grâce à une méthode de *Shift Invariant Sparse Coding*.

Table 2. Tableau récapitulatif des modalités utilisées dans les précédents travaux

Travaux	Temporalité	Prosodie	Vidéo	Texte
[34]	Non	Oui	Oui	Non

[30]	Non	Oui	Oui	Non
[32]	Non	Oui	Oui	Oui
[6]	Locale	Oui	Oui	Oui
[44]	Non	Oui	Oui	Oui
[48]	Non	Oui	Oui	Non
[7]	Locale	Oui	Oui	Oui
[14]	Oui	Oui	Oui	Non
Cette étude	Non	Non	Non	Oui

3. CONSTITUTION DU CORPUS ET ANNOTATIONS

Nous utilisons un corpus de données fourni par l'entreprise EASYRECRUE. Cette société propose un service d'entretien vidéo différé. Le corpus compte 607 candidats français ayant chacun répondu à six questions en vidéo et à six questions à l'écrit pour un poste de conseiller commercial. L'entretien est structuré, les questions ont été choisies par le recruteur. Parmi les questions à réponse vidéo, on peut retrouver des questions interrogeant sur l'expérience du candidat, sa motivation, des questions situationnelles et techniques. Les réponses à l'écrit correspondent à des questions concernant la disponibilité, le dernier diplôme, le temps de trajet pour le bureau ou la rémunération demandée. Le temps de préparation pour chaque candidat est de 30s et le temps de réponse accordé varie entre 40 et 90s. Un seul et même recruteur a évalué ces candidats dans le cadre de son activité. Lorsque les candidats ont complété leur entretien, ce recruteur le regarde et peut choisir de cliquer sur un bouton « avis favorable » ou « avis réservé ». Il est libre de noter, regarder en totalité ou non les réponses des candidats.

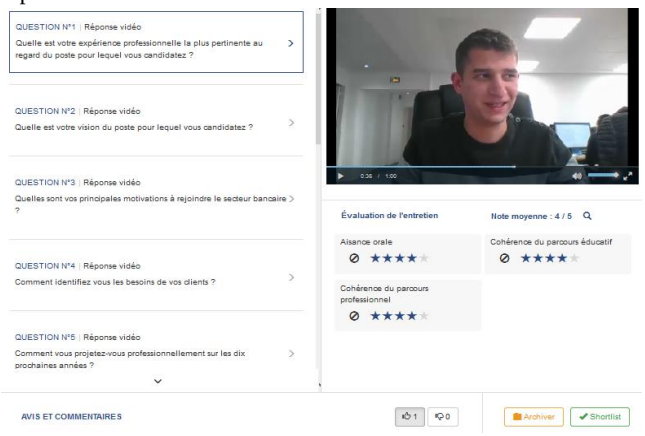

Figure 1. Interface d'évaluation du recruteur

Dans un premier temps, seules les transcriptions automatiques des questions vidéo sont étudiées. Ces transcriptions ont été effectuées grâce à l'*Application Programming Interface* (API) de reconnaissance automatique de la parole de Google[1]. La qualité de celle-ci dépend largement de la piste audio enregistrée. Après une étude qualitative, nous avons décidé de filtrer les candidats pour lesquels, la moitié des transcriptions ou plus comportaient moins de dix mots, car la majorité de ces transcriptions étaient très mauvaises. Nous envisageons dans nos prochains travaux d'utiliser les scores de confiance associés aux transcriptions pour repérer les transcriptions peu fiables (pour cette première étude, nous ne disposions pas encore de ces scores). Seuls les candidats évalués « avis favorable » ou « avis réservé » sont utilisés dans cette étude. Après application des critères de sélection évoqués, le nombre de candidats restant est égal à 305. Plusieurs statistiques décrivant le corpus ainsi constitué sont disponibles dans la table 3.

Table 3. Tableau descriptif du corpus constitué

Nombre de candidats	305
Pourcentage de candidats « favorable »	0.61
Nombre moyen de mots d'une réponse à une question	77 mots
Nombre de mots total	139605
Nombre de mots uniques	8067

4. ANALYSE PRÉDICTIVE

L'objectif de cette partie est de proposer un modèle prédictif afin de trier automatiquement des candidats. La tâche est une classification binaire entre candidats étiquetés « avis favorable » ou « avis réservé ». Les candidats sont ensuite classés en fonction du score de confiance attribué par le modèle prédictif. Tout d'abord, nous présentons les descripteurs extraits du contenu verbal du candidat puis le modèle et le protocole d'évaluation seront expliqués.

4.1 Extraction des descripteurs

La psychologie sociale joue un rôle important dans le domaine de l'analyse des entretiens d'embauche. Afin de modéliser les informations liées aux comportements socio-émotionnels - et à leur dimension psychologique - présentes dans le contenu verbal, nous avons choisi d'utiliser deux dictionnaires : LIWC traduit en français par A.Piolat et Al [41] et FEEL [42]. Le dictionnaire LIWC recense à la fois des mots correspondant à une catégorie d'émotions et des catégories grammaticales. Les catégories de mots sont organisées selon les différents processus psychologiques liés aux émotions. [41] . Ce dictionnaire va nous permettre d'extraire des informations du type : l'utilisation du « je » ou du « nous » influence-t-il l'évaluation du recruteur ? Qu'en est-il de l'utilisation de mots reliés aux notions de perspicacité ou de colère ? Le dictionnaire FEEL [42] référence plus de 14 000 mots distincts en fonction de leur polarité et de leur appartenance à l'une des 6 émotions basiques définies par Ekman. Par exemple, l'utilisation de ce dictionnaire nous semble intéressante pour caractériser si la présence de mots négatifs et de mots évoquant un sentiment de peur a une influence sur la perception que le recruteur a du candidat. La relation entre **la diversité et la singularité lexicale** et l'évaluation d'un recruteur nous semble être une piste à explorer. Des statistiques peuvent nous renseigner sur la diversité lexicale et sont apparues efficaces [32,

[1] https://cloud.google.com/speech/

45]. Afin de capturer cette diversité, nous avons choisi d'extraire quatre indices - le *Type token Ratio*, l'indice HD-D [28], *the measure of textual lexical diversity* MTLD [28] et l'indice de lecture de Kandel-Moles [23], - la densité du nombre de mots de plus de 6 lettres et la longueur moyenne des mots. De plus, il a été prouvé que le débit de parole influence grandement l'évaluation du recruteur [17, 32]. Ainsi, nous capturons le nombre de mots de la réponse du candidat normalisé par la réponse la plus longue. Nous pensons que l'utilisation des mots de liaison structure le monologue et peut-être une information importante quant à la qualité du contenu verbal. Un dictionnaire de mots de liaison a donc été construit en nous appuyant sur différentes pages[2][3]. La singularité lexicale se mesure dans le fait qu'un discours utilise des mots plus ou moins rares. Dans le but de mesurer celle-ci, nous utilisons la base de données Lexique3 [33]. Cette base référence plus de 135 000 mots français et fournit des informations pour chacun de ces mots telles que la fréquence d'utilisation dans des corpus de films ou de livres. Dans la continuité, **le registre de langue** semble aussi intéressant à évaluer, un registre plus ou moins familier ou soutenu pourrait influencer l'évaluation d'un recruteur. Un dictionnaire a été construit en se basant sur la page de Wiktionary[4] relative au registre afin de répondre à ce besoin. La construction des descripteurs des dictionnaires LIWC, FEEL, registre et mots de liaison s'opère en comptant les mots de chaque dimension des dictionnaires, ce compte est ensuite normalisé par le nombre de mots de la réponse. De plus, les catégories représentées dans moins de 10% des réponses sont évincées. Pour les descripteurs issus de Lexique3, nous extrayons pour chacun des mots la fréquence de son utilisation dans un corpus de films et de livres, le nombre de syllabes et le pourcentage de personnes connaissant le mot puis nous agrégeons les valeurs au niveau de la réponse en utilisant la moyenne, l'écart type, le 1er et le 3e quartile. **La densité grammaticale** a été analysée dans plusieurs travaux antérieurs [35, 43]. Il nous semble important d'analyser si l'utilisation de catégories grammaticales telle que les adjectifs ou les adverbes peut avoir une influence sur l'évaluation. Ainsi, pour chaque mot, sa catégorie grammaticale est détectée grâce à l'outil TreeTagger [50]. La densité d'utilisation de chaque catégorie grammaticale est ensuite calculée. Enfin dans le but de représenter au mieux **les mots contenus dans la réponse et la sémantique employée**, nous utilisons un algorithme de représentation nommé Doc2Vec implémenté sous gensim [47]. Doc2Vec a été utilisé efficacement dans plusieurs études telles que l'évaluation automatique de candidats [6]. Doc2Vec est une méthode non supervisée qui consiste à projeter des textes dans un espace sémantique où les textes les plus similaires tendent à être plus proches les uns des autres. Nous avons entraîné Doc2Vec sur un ensemble de transcriptions provenant d'entretiens d'embauche différés afin de projeter efficacement les réponses des différents candidats dans un espace de dimension 100. Les différents blocs de descripteurs sont résumés dans la table 4.

Table 4. Tableau descriptif des différents blocs de descripteurs

Descripteurs	Dimension	Exemples	Réfs
LIWC	58	Utilisation du « je », utilisation du « nous », mots remplisseurs, utilisation de mots appartenant à la catégorie perspicacité, …	[37, 49, 57]
Registre	11	Termes familiers, argot en français, termes littéraires, termes péjoratifs, termes populaires, …	[46]
Diversité lexicale et structure	8	Mots de plus de 6 lettres, HD-D, MTLD, Mots de liaison, longueur de la réponse normalisée, …	[8, 28, 40, 49, 58]
Nature des mots employés	10	Nom propre, verbe, adverbe, nom, adjectif, …	[40, 44, 50]
Lexique3	42	Fréquences d'apparition dans la langue française écrite et orale, nombre de syllabes, difficulté apparente des mots, …	[33]
FEEL	7	Polarité des mots et appartenance à l'une des 6 dimensions définies par Ekman	[42]
Doc2Vec	100	Descripteurs issus de l'extraction par l'algorithme Doc2Vec	[6, 25]

4.2 MODELE ET PROTOCOLE D'ÉVALUATION

Nous construisons un modèle de classification binaire selon les préférences d'un seul et même recruteur entre les candidats étiquetés « avis favorable » ou « avis réservé ». Une donnée en entrée du classifieur correspond à la réponse à une question pour un candidat soit 1796 réponses. Afin de mesurer la performance du modèle, nous utilisons l'aire sous la courbe ROC [19]. Pour évaluer notre méthode de classification, nous séparons le jeu de données en un jeu d'entraînement et un jeu de test représentant respectivement 5/6 et 1/6 du jeu de données total. Nous entraînons notre modèle de classification binaire sur le jeu d'entraînement après avoir choisi les hyper paramètres en effectuant une validation croisée à 5 partitions. Nous évaluons ensuite notre modèle sur le jeu de test. Ce

[2]http://www.mycampus-live.com/telechargements/4-B2-argumentation-Tableaux-mots-de-liaison-et-modalisateurs.

[3]http://hyperpolyglotte.com/apprends-francais/vocabulaire-connecteurs.php

[4]https://fr.wiktionary.org/wiki/Cat%C3%A9gorie:Registres_de_langue_en_fran%C3%A7ais

processus est itéré 300 fois afin d'éviter tout artefact présent lors du découpage des données. Lorsque l'on constitue le jeu d'entraînement et les partitions de la validation croisée, nous veillons à ce qu'un candidat ne se retrouve que dans une seule partition. Les classifieurs testés sont la régression logistique régularisée par la méthode Lasso et les forêts aléatoires. Les hyper paramètres pour le Lasso et les forêts aléatoires sont choisis automatiquement lors de la validation croisée. La valeur moyenne obtenue pour chacune des catégories de descripteurs est évaluée selon cette méthode. Aussi, deux types de fusions sont essayés : la fusion précoce (early fusion) et la fusion tardive (late fusion). La fusion précoce consiste à utiliser tous les descripteurs sans distinction de catégories ; la fusion tardive, à moyenner les probabilités de sorties de chaque catégorie.

5. RÉSULTATS

Dans cette section, nous présentons les résultats obtenus. Le Tableau 5 présente les résultats obtenus par les différents classifieurs en fonction des catégories de descripteurs choisis. Le meilleur résultat est obtenu avec les forêts aléatoires en utilisant la stratégie de fusion tardive entre les différents blocs de descripteurs. Le score d'aire sous la courbe ROC (0.69) est bien supérieur à celui qui serait obtenu avec l'aléatoire (0.5). La diversité et la singularité lexicale et la sémantique semblent mieux discriminer que le caractère psychologique, le registre utilisé ou les émotions véhiculées. Le modèle employé semble aussi influencer la note obtenue spécifiquement pour le modèle Lasso. Plusieurs hypothèses sont possibles afin d'expliquer la différence du score obtenu entre le modèle Lasso et les forêts aléatoires : notamment des relations non linéaires entre descripteurs ou une pénalisation trop importante de descripteurs par le modèle Lasso. Une exploration des données a été effectuée après la construction de nos modèles. Ainsi, pour le modèle prenant en entrée tous les descripteurs, nous inspectons le pourcentage de sélection de chacun d'eux par le modèle Lasso au cours des 300 itérations. Les trois descripteurs les plus sélectionnés sont liés à l'indice de lecture, la longueur des réponses aux questions et la singularité lexicale moyenne. De plus, nous pouvons également remarquer que la densité des mots de liaisons est sélectionnée comme étant un descripteur discriminant. L'utilisation des mots de liaisons informe sur le niveau de structure des réponses. Ainsi, il semble qu'une réponse structurée ait une meilleure évaluation. Enfin, les descripteurs provenant des dictionnaires FEEL et Registre jouent un rôle négligeable dans la classification alors que les descripteurs issus du LIWC occupent une part plus importante dans la classification (sept descripteurs LIWC sélectionnés dans les 40 premiers).

Table 5. Tableau des résultats obtenus

Bloc de descripteurs	AUC			
	Forêts aléatoires		Lasso	
	Moyenne	Écart type	Moyenne	Écart type
LIWC	0.640	0.074	0.569	0.088
Registre	0.638	0.073	0.500	0.079
Diversité lexicale et structure	0.691	0.079	**0.691**	0.078
Nature des mots employés	0.650	0.077	0.490	0.069
Lexique 3	0.685	0.070	0.687	0.071
FEEL	0.662	0.073	0.572	0.074
Doc2Vec	0.670	0.069	0.668	0.072
Tous les descripteurs	0.691	0.075	0.689	0.076
Fusion tardive des blocs des descripteurs	**0.696**	0.073	0.695	0.076

6. CONCLUSION ET PERSPECTIVES

Dans cet article, nous avons présenté une analyse automatique de candidats ayant passé des entretiens vidéo différés basés sur leur contenu verbal. Après collecte d'un jeu de données en situation réelle d'entretien, nous avons montré la faisabilité de la construction d'un modèle de classement automatique. Après analyse de ce modèle, les candidats fournissant des réponses plus longues, utilisant fréquemment des mots de liaison, une grande diversité et une grande singularité lexicale sont mieux classés. Les émotions véhiculées par le candidat au travers de son discours et le registre de langue employé ne semblent pas influencer l'évaluateur, du moins linéairement. Nous envisageons plusieurs suites à ces travaux. Premièrement, nous nous sommes concentrés pour cette première étude sur le contenu verbal, mais nous envisageons d'étudier la complémentarité des modalités notamment pour l'analyse des comportements émotionnels. En effet, nous pensons que les premières impressions vis-à-vis de l'état d'un candidat (stressé, apeuré, confiant, …) sont véhiculées également par sa prosodie (ex : débit de parole), des phénomènes disfluents (hésitations, répétitions) qui n'apparaissent pas dans la transcription (et peuvent même entraver celle-ci [10]) et ses comportements non verbaux (expressions faciales, posture). Deuxièmement, cette première étude s'appuie sur l'évaluation réalisée par un unique recruteur. Nous souhaitons étendre l'étude à plusieurs recruteurs. Troisièmement, cette étude se focalise sur la classification binaire obtenue avec l'évaluation finale (« avis favorable » ou « avis réservé »). Nous souhaitons analyser les critères intermédiaires (ex: motivation, originalité, perspicacité, ...) utilisés par les recruteurs pour leur évaluation finale. Enfin, les méthodes d'apprentissage utilisées dans cette première étude ne prennent pas en compte la dynamique temporelle des signaux sociaux (comportements verbaux et non verbaux), dynamique qui joue un rôle primordial dans l'expression des comportements socio-émotionnels. Des méthodes de fouille de séquences [22, 55] sont envisagées afin de trouver des séquences de comportements verbaux et non verbaux influençant l'évaluation des recruteurs.

7. RÉFÉRENCES

[1] Ali, M.R. et al. 2015. LISSA - Live Interactive Social Skill Assistance. *2015 International Conference on Affective Computing and Intelligent Interaction, ACII 2015*. (2015), 173-179. DOI:https://doi.org/10.1109/ACII.2015.7344568.

[2] Anderson, K. et al. 2013. The TARDIS framework : intelligent virtual agents for social coaching in job interviews. (2013).

[3] Batrinca, L. et al. 2013. Cicero - Towards a multimodal virtual audience platform for public speaking training. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes*

in Bioinformatics). 8108 LNAI, (2013), 116-128. DOI:https://doi.org/10.1007/978-3-642-40415-3_10.

[4] Biel, J.-I. et Gatica-Perez, D. 2013. The YouTube Lens: Crowdsourced Personality Impressions and Audiovisual Analysis of Vlogs. *IEEE Transactions on Multimedia*. 15, 1 (janv. 2013), 41-55. DOI:https://doi.org/10.1109/TMM.2012.2225032.

[5] Booth, B.M. et al. 2017. Toward Active and Unobtrusive Engagement Assessment of Distance Learners. (2017), 470-476.

[6] Chen, L. et al. 2016. Automated scoring of interview videos using Doc2Vec multimodal feature extraction paradigm. *Proceedings of the 18th ACM International Conference on Multimodal Interaction - ICMI 2016*. October (2016), 161-168. DOI:https://doi.org/10.1145/2993148.2993203.

[7] Chen, L. et al. Automated Video Interview Judgment on a Large-Sized Corpus Collected Online.

[8] Chen, L. et al. 2009. Improved Pronunciation Features for Construct-driven Assessment of Non-native Spontaneous Speech. June (2009), 442-449.

[9] Chen, L. et al. 2014. Towards Automated Assessment of Public Speaking Skills Using Multimodal Cues. *Icmi 2014*. September (2014), 1-4. DOI:https://doi.org/10.1145/2663204.2663265.

[10] Clavel, C. et al. 2013. *Spontaneous speech and opinion detection: Mining call-centre transcripts*.

[11] Conway, J.M. et al. 1995. A meta-analysis of interrater and internal consistency reliability of selection interviews. *Journal of Applied Psychology*. 80, 5 (1995), 565-579. DOI:https://doi.org/10.1037/0021-9010.80.5.565.

[12] Damian, I. et al. 2013. Investigating social cue-based interaction in digital learning games. *... of the 8th International Conference on ...*. (2013).

[13] Escalante, H.J. et al. 2017. ChaLearn Joint Contest on Multimedia Challenges beyond Visual Analysis: An overview. *Proceedings - International Conference on Pattern Recognition*. (2017), 67-73. DOI:https://doi.org/10.1109/ICPR.2016.7899609.

[14] Escalante, H.J. et al. Visualizing Apparent Personality Analysis with Deep Residual Networks. 3101-3109.

[15] Finnerty, A.N. et al. 2016. Stressful first impressions in job interviews. *Proceedings of the 18th ACM International Conference on Multimodal Interaction - ICMI 2016*. October (2016), 325-332. DOI:https://doi.org/10.1145/2993148.2993198.

[16] G, D.A. et al. 2017. Toward Automatic Detection of Acute Stress : Relevant Nonverbal Behaviors and Impact of Personality Traits. (2017), 354-361.

[17] Gifford, R. et al. 1985. Nonverbal cues in the employment interview: Links between applicant qualities and interviewer judgments. *Journal of Applied Psychology*. 70, 4 (1985), 729-736. DOI:https://doi.org/10.1037/0021-9010.70.4.729.

[18] Güçlütürk, Y. et al. 2016. Deep Impression: Audiovisual Deep Residual Networks for Multimodal Apparent Personality Trait Recognition. September (2016). DOI:https://doi.org/10.1007/978-3-319-49409-8_28.

[19] Hanley, A.J. et McNeil, J.B. 1982. The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve. *Radiology*. 143, (1982), 29-36. DOI:https://doi.org/10.1148/radiology.143.1.7063747.

[20] Hartwell, C.J. et al. 2014. THE STRUCTURED EMPLOYMENT INTERVIEW : NARRATIVE AND QUANTITATIVE REVIEW OF THE RESEARCH LITERATURE. (2014), 241-293. DOI:https://doi.org/10.1111/peps.12052.

[21] Hoque, M.E. et al. 2016. Mach: My automated conversation coach. *UbiComp '13: Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*. (2016), 697-706. DOI:https://doi.org/10.1145/2493432.2493502.

[22] Janssoone, T. et al. SMART : Règles d ' associations temporelles de signaux sociaux pour la synthèse d ' un Agent Conversationnel Animé avec une attitude spécifique. 1-16. DOI:https://doi.org/10.3166/RIA.28.1-.

[23] KANDEL, L. et MOLES, A. 1958. Application de l'indice de flesch à la langue française. *Cahiers Etudes de Radio-Télévision*. 19, (1958), 253-274.

[24] Kaya, H. et al. 2017. Video-based emotion recognition in the wild using deep transfer learning and score fusion. *Image and Vision Computing*. 65, (2017), 66-75. DOI:https://doi.org/10.1016/j.imavis.2017.01.012.

[25] Le, Q. et al. 2014. Distributed Representations of Sentences and Documents. 32, (2014).

[26] Li, R. et al. 2018. Understanding Social Interpersonal Interaction via Synchronization Templates of Facial Events. *The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*. (2018).

[27] Macan, T. 2009. The employment interview: A review of current studies and directions for future research. *Human Resource Management Review*. 19, 3 (2009), 203-218. DOI:https://doi.org/10.1016/j.hrmr.2009.03.006.

[28] McCarthy, P.M. et Jarvis, S. 2010. MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*. 42, 2 (2010), 381-392. DOI:https://doi.org/10.3758/BRM.42.2.381.

[29] Muralidhar, S. et al. 2017. How may I help you? behavior and impressions in hospitality service encounters. *Proceedings of the 19th ACM International Conference on Multimodal Interaction - ICMI 2017*. (2017), 312-320. DOI:https://doi.org/10.1145/3136755.3136771.

[30] Muralidhar, S. et al. 2016. Training on the job: behavioral analysis of job interviews in hospitality. *Proceedings of the 18th ACM International Conference on Multimodal Interaction - ICMI 2016*. (2016), 84-91. DOI:https://doi.org/10.1145/2993148.2993191.

[31] Muralidhar, S. et Gatica-perez, D. 2017. Examining Linguistic Content and Skill Impression Structure for Job Interview Analytics in Hospitality. (2017), 339-343.

[32] Naim, I. et al. 2015. Automated prediction and analysis of job interview performance: The role of what you say and how you say it. *2015 11th IEEE International Conference*

and Workshops on Automatic Face and Gesture Recognition, FG 2015 (2015).

[33] New, B. et al. Une base de données lexicales du français contemporain sur internet : LEXIQUE ™. 1-21.

[34] Nguyen, L.S. et al. 2014. Hire me: Computational inference of hirability in employment interviews based on nonverbal behavior. *IEEE Transactions on Multimedia*. 16, 4 (2014), 1018-1031. DOI:https://doi.org/10.1109/TMM.2014.2307169.

[35] Nguyen, L.S. et Gatica-Perez, D. 2016. Hirability in the Wild: Analysis of Online Conversational Video Resumes. *IEEE Transactions on Multimedia*. 18, 7 (2016), 1422-1437. DOI:https://doi.org/10.1109/TMM.2016.2557058.

[36] Nguyen, L.S.D.G.-P. 2015. I Would Hire You in a Minute: Thin Slices of Nonverbal Behavior in Job Interviews. *Icmi*. (2015). DOI:https://doi.org/10.1145/2818346.2820760.

[37] Niederhoffer, K.G. et Pennebaker, J.W. 2002. LINGUISTIC STYLE MATCHING IN SOCIAL INTERACTION. 21, 4 (2002), 337-360. DOI:https://doi.org/10.1177/026192702237953.

[38] Paper, C. et Nam, H.K. 2017. Multi-modal Score Fusion and Decision Trees for Explainable Automatic Job Candidate Screening from Video CVs. October (2017). DOI:https://doi.org/10.1109/CVPRW.2017.210.

[39] Peck, J.A. et Levashina, J. 2017. Impression management and interview and job performance ratings: A meta-analysis of research design with tactics in mind. *Frontiers in Psychology*. 8, FEB (2017), 1-10. DOI:https://doi.org/10.3389/fpsyg.2017.00201.

[40] Persing, I. et Ng, V. Modeling Argument Strength in Student Essays.

[41] Piolat, A. et al. 2011. La version franaise du dictionnaire pour le LIWC: modalités de construction et exemples d'utilisation. *Psychologie Francaise*. 56, 3 (2011), 145-159. DOI:https://doi.org/10.1016/j.psfr.2011.07.002.

[42] Poncelet, P. 2016. FEEL : a French Expanded Emotion Lexicon. (2016). DOI:https://doi.org/10.1007/s10579-016-9364-5.

[43] Rasipuram, S. et al. 2017. Automatic prediction of fluency in interface-based interviews. *2016 IEEE Annual India Conference, INDICON 2016*. December (2017). DOI:https://doi.org/10.1109/INDICON.2016.7838991.

[44] Rasipuram, S. Prediction/Assessment of Communication Skill using Multimodal Cues in Social Interactions.

[45] Rasipuram, S. et Jayagopi, D.B. 2016. Automatic assessment of communication skill in interface-based employment interviews using audio-visual cues. *2016 IEEE International Conference on Multimedia and Expo Workshop, ICMEW 2016*. September (2016). DOI:https://doi.org/10.1109/ICMEW.2016.7574733.

[46] Registres de langue en français: *https://fr.wiktionary.org/wiki/Catégorie:Registres_de_langue_en_français*.

[47] Rehurek, R. et Sojka, P. 2010. Software Framework for Topic Modelling with Large Corpora. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. (2010), 45-50. DOI:https://doi.org/10.13140/2.1.2393.1847.

[48] Rupasinghe, A.T. et al. 2017. Scaling personality traits of interviewees in an online job interview by vocal spectrum and facial cue analysis. *16th International Conference on Advances in ICT for Emerging Regions, ICTer 2016 - Conference Proceedings*. September (2017), 288-295. DOI:https://doi.org/10.1109/ICTER.2016.7829933.

[49] Sanchez-cortes, D. et al. 2012. Assessing the Impact of Language Style on Emergent Leadership Perception from Ubiquitous Audio. (2012).

[50] Schmid, H. (IMS-C. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. (1994).

[51] Schmidt, F.L. et Hunter, J.E. 1998. The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*. 124, 2 (1998), 262-274. DOI:https://doi.org/10.1037/0033-2909.124.2.262.

[52] Schmidt, G.F. 2007. The effect of thin slicing on structured interview decisions. (2007).

[53] Straus, S.G. et al. 2001. The effects of videoconference, telephone, and face-to-face media on interviewer and applicant judgments in employment interviews. *Journal of Management*. 27, 3 (2001), 363-381. DOI:https://doi.org/10.1016/S0149-2063(01)00096-4.

[54] Tanaka, H. et al. 2015. Automated Social Skills Trainer. *Proceedings of the 20th International Conference on Intelligent User Interfaces*. (2015), 17-27. DOI:https://doi.org/10.1145/2678025.2701368.

[55] Tanveer, M.I. et al. 2016. AutoManner: An Automated Interface for Making Public Speakers Aware of Their Mannerisms. *Proceedings of the 21st International Conference on Intelligent User Interfaces*. (2016), 385-396. DOI:https://doi.org/10.1145/2856767.2856785.

[56] Tanveer, M.I. et Lin, E. Rhema : A Real-Time In-Situ Intelligent Interface to Help People with Public Speaking.

[57] Walker, M.A. et al. 2007. Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text. 30, (2007), 457-500.

[58] Zechner, K. et Bejar, I.I. 2006. Towards Automatic Scoring of Non-Native Spontaneous Speech. section 7 (2006), 1-8.

[59] Zhao, R.U. et al. 2017. Semi-Automated & Collaborative Online Training Module for Improving Communication Skills. 1, 2 (2017), 1-20.