Attention Slices dans les Entretiens d'Embauche Vidéo Différés

Léo Hemamou, Arthur Guillon, Jean-Claude Martin, Chloé Clavel

Attention Slices dans les Entretiens d'Embauche Vidéo Différés

Léo Hemamou[*†]
EASYRECRUE
Paris, France
l.hemamou@easyrecrue.com

Arthur Guillon
EASYRECRUE
Paris, France
a.guillon@easyrecrue.com

Jean-Claude Martin
LIMSI, CNRS, Paris-Sud University, Paris-Saclay University, Orsay, France
Jean-Claude.Martin@limsi.fr

Chloé Clavel
LTCI, Télécom ParisTech, Paris-Saclay University
Palaiseau, France
chloe.clavel@telecom-paristech.fr

ABSTRACT

Dans cet article, nous nous intéressons à l'étude de signaux influents dans les entretiens vidéo d'embauche asynchrones découverts par des méthodes d'apprentissage profond. Le système que nous étudions emploie des mécanismes d'attention, qui permettent d'extraire d'un entretien les informations et les instants décisifs (qui ont influencé la décision du système au niveau de l'entretien), sans requérir d'annotation locale. Alors que la majorité des approches similaires évaluent les mécanismes d'attention en se contentant de visualiser les moments d'attention maximale, nous proposons ici une méthodologie permettant d'automatiser l'analyse du contenu de ces *attention slices* afin de fournir des éléments d'interprétation des prédictions du système.

KEYWORDS

entretiens d'embauche, entretiens vidéo, apprentissage profond, interprétabilité

ACM Reference Format:
Léo Hemamou, Arthur Guillon, Jean-Claude Martin, and Chloé Clavel. 2020. *Attention Slices* dans les Entretiens d'Embauche Vidéo Différés. In *WACAI 2020, 03–05 Juin, 2020, Île d'Oléron, France.* ACM, New York, NY, USA, 8 pages. https://doi.org/https://doi.org/xx.xxxx/xxxxxxx.xxxxxxx

1 INTRODUCTION

Parmi les procédures de sélection du personnel, les entretiens d'embauche demeurent une méthode privilégiée des recruteurs [29]. L'informatique affective a déjà prouvé, à de nombreux égards, son utilité dans ce contexte : par exemple, des recruteurs virtuels ont été pensés pour aider les candidats à développer leurs compétences sociales et à répéter l'exercice de l'entretien d'embauche [16]. Dès lors, l'apport du traitement automatique peut être bénéfique à l'entraînement des candidats mais aussi aux chercheurs et aux recruteurs pour comprendre le processus d'évaluation et de sélection opéré par ces derniers. Initialement menés en face-à-face ou par téléphone, les entretiens d'embauche sont désormais souvent réalisés par des systèmes de visioconférence en ligne ou par des enregistrements vidéo asynchrones. La procédure est la suivante : le candidat se connecte à une plateforme web et répond à une séquence de questions prédéfinies par le recruteur en s'enregistrant en vidéo au moyen de sa webcam, son smartphone ou sa tablette. Dans un second temps, les recruteurs se connectent à la même plateforme, regardent les réponses du candidat, évaluent les réponses et décident ensuite s'ils souhaitent inviter le candidat à un entretien en face à face.

Les chercheurs développent déjà des systèmes pour prédire automatiquement l'embauche sur la base d'indices non verbaux des candidats dans des interviews vidéo asynchrones [4, 27]. Dans ce contexte et en plus des nouvelles contraintes législatives (Règlement Général sur la Protection des Données), de tels systèmes automatiques nécessitent de l'interprétabilité et de la transparence. Grâce à ces systèmes, les candidats pourront améliorer leur stratégie comportementale non verbale et les recruteurs pourront évaluer ces modèles d'aide à la décision. Ces modèles pourraient même aider les recruteurs à comprendre leurs propres biais. Nous avons précédemment proposé un modèle d'apprentissage profond, baptisé HireNet, entraîné uniquement grâce à l'étiquette de la décision du recruteur et des vidéos réponses d'entretiens d'embauche [14]. Notre modèle est capable de prendre en compte la temporalité et les tranches influentes des entretiens vidéos différés, grâce à l'utilisation d'un mécanisme d'attention et d'un réseau de neurones récurrent. Une séquence de descripteurs est traitée par un réseau de neurones récurrent, et le mécanisme d'attention vise à apprendre un poids différent pour chaque élément (ou pas de temps) de cette séquence afin d'améliorer les performances de la tâche de classification. Dans l'ensemble, ces techniques pourraient être utiles pour comprendre le comportement humain, car elles visent à séparer les pas de temps pertinents pour une tâche dans une séquence des pas de temps non pertinents [33]. Cependant, la recherche dans ce domaine se limite à montrer quelques exemples de pics de courbes d'attention [14, 33]. Par conséquent, une validation cohérente est nécessaire afin de vérifier l'utilité de la sortie d'un tel système pour prédire l'embauche telle qu'évaluée par les recruteurs.

Dans cet esprit, cet article décrit quatre expériences que nous avons menées pour comprendre si des tranches d'entretiens vidéo

[*]Also with LIMSI, CNRS, Paris-Sud University, Paris-Saclay University.
[†]Also with LTCI, Télécom ParisTech, Paris-Saclay University.

WACAI 2020, 03–05 Juin, 2020, Île d'Oléron, France

ACM ISBN XXX-X-XXXX-XXXX-X/XX/XX…$15.00
https://doi.org/https://doi.org/xx.xxxx/xxxxxxx.xxxxxxx

mises en évidence par le modèle d'attention contiennent des informations utiles aux recruteurs. Tout d'abord en section 2, nous présentons un état de l'art des méthodes d'analyse automatique d'entretien d'embauche. La section 3 décrit le nouveau jeu de données collectées et rappelle le modèle d'apprentissage profond que nous utilisons, HireNet, puis le compare à des modèles de l'état de l'art. Enfin la section 4 propose une étude approfondie des courbes d'attention produites par HireNet.

2 ÉTAT DE L'ART

2.1 L'influence des comportements verbaux et non verbaux dans le contexte du recrutement

Comportement non verbaux dans les entretiens. Les comportements non verbaux tels que les indices visuels et audio ont été étudiés par plusieurs auteurs dans le contexte du recrutement, notamment, en lien avec la prédiction de l'anxiété [11], des performances lors de l'entretien [12], de la personnalité des candidats [7] ou des tentatives de tromperie [30]. De nombreuses caractéristiques visuelles ont ainsi été étudiées, telles que l'attractivité du candidat, la gestuelle des mains, le sourire, le contact visuel, le hochement ou le mouvement de la tête, l'orientation du corps ou les expressions faciales [11, 30]. Des travaux comparables ont été menés pour la voix, en particulier concernant l'impact de l'amplitude et du ton de la voix, les silences ou les dysfluences verbables [24, 27].

Contenu verbal dans les entretiens. En comparaison avec ces travaux, peu d'études se sont focalisées sur l'influence du contenu verbal [22]. Celui-ci est pourtant central dans l'évaluation de la performance du candidat, puisqu'il est la modalité utilisée par le candidat pour répondre. En tant que tel, le contenu verbal constitue une grande source d'informations pour la plupart des dimensions évaluées lors d'un entretien d'embauche telles que les connaissances déclaratives liées au poste, les compétences sociales, communicationnelles ou interpersonnelles [17].

L'annotation de chaque comportement du candidat, qu'il soit verbal ou non verbal, est coûteuse en temps. Une approche courante pour alléger la tâche d'annotation consiste à annoter uniquement une partie de l'entretien d'embauche. En fait, il a été démontré que, en utilisant seulement une petite quantité d'informations, les gens peuvent déduire correctement les caractéristiques, traits ou états personnels d'un individu [2, 23]. Cette approche est appelée analyse en tranches fines (*thin slices*) et a déjà été utilisée dans l'étude des interactions sociales [23], les premières impressions [2], la prise de parole en public [6] ou les entretiens d'embauche [25]. Un autre avantage de cette méthode est qu'elle met en évidence un comportement bref et non verbal par rapport aux impressions perçues. Néanmoins, la durée et la stratégie d'échantillonnage des tranches fines restent une question ouverte. Les études précédentes se concentrent sur l'échantillonnage de tranches fines au hasard [6], en utilisant la structure de l'entretien d'embauche (tranches basées sur des questions et réponses) [25], ou au début et à la fin des interactions [7]. Des méthodes automatiques basées sur le traitement du signal social pourraient laisser la place à une meilleure sélection des tranches fines et de leur durée, en sélectionnant des régions qui véhiculent plus d'informations.

2.2 Analyse automatique des entretiens d'embauche

Les avancées récentes réduisent considérablement le temps passé à annoter manuellement les signaux comportementaux. Des outils sont désormais disponibles pour détecter automatiquement les signaux vocaux [9] ou visuels [1]. Des études récentes utilisent le traitement du signal social et l'apprentissage automatique pour comprendre les liens entre les indices non verbaux et l'embauche. Ces études ont été appliquées à différentes déclinaisons de l' entretien d'embauche : en entretiens en face à face [24, 25], en entretiens vidéo asynchrones [4] ou entretiens par chat vidéo [26].

Parmi les dimensions étudiées dans les entretiens d'embauche (compétences communicationnelles [26], personnalité [4], etc.), la performance en entretien reste la plus étudiée. La méthodologie la plus usitée est l'extraction de descripteurs selon chacune des modalités, l'obtention d'un vecteur fixe par application de fonctionnelles (moyenne, écart type, minimum ou maximum des valeurs des descripteurs au cours d'une réponse) puis l'entraînement d'un modèle d'apprentissage automatique. Cependant, cette méthode présente des lacunes, notamment l'absence de la modélisation du caractère séquentiel de l'entretien et son aspect hiérarchique : un entretien d'embauche est une séquence de questions/réponses dont chacune des réponses est elle-même une séquence de mots ou de fenêtres. Pour répondre à ce manque, nous avons proposé un modèle d'apprentissage profond [14] ayant démontré des meilleurs résultats sur la tâche de classification d'employabilité. Cependant, cette méthode étant basée sur l'apprentissage profond, le système d'apprentissage est bien plus opaque qu'un système basé sur l'apprentissage automatique et des descripteurs construits à la main. Il est donc nécessaire d'effectuer une analyse approfondie du système afin de comprendre comment celui-ci fonctionne et nous informer sur des possibles comportements intéressants à étudier dans le contexte du recrutement.

2.3 Réseaux de neurones et interprétabilité

Les réseaux de neurones sont capables de modéliser une représentation intermédiaire intégrant une information haut niveau directement à partir des données signal [31]. Cependant, la liberté dont ils disposent pour construire cette représentation intermédiaire se fait au prix d'une extrême opacité qui freine leur adoption pour des applications critiques telles que dans le milieu de la santé, de la justice ou des ressources humaines. Pour ces raisons, de nombreux travaux proposent des méthodes pour mieux expliquer ces réseaux. La notion de transparence et d'interprétabilité reste encore floue, notamment dans le cas où la vérité terrain n'est pas connue [32]. [32] propose une catégorisation à deux dimensions pour l'interprétabilité selon la dimension de l'interprétation (explique-t-on le modèle dans son ensemble, ou la prédiction sur une instance ?) et la méthode d'interprétation (intrinsèque au modèle, ou par une méthode post-construction du modèle ?). Plusieurs méthodes ont ainsi été proposées comme la visualisation des états cachés, les approches par compression de modèle [20], l'utilisation de classifieurs locaux [28]. Parmi ces méthodes, les mécanismes d'attention ont récemment gagné en popularité pour améliorer les performances et l'interprétabilité [14, 33]. Cependant, la plupart des études existantes se contentent de valider la pertinence des mécanismes

d'attention en exhibant des exemples et ne conduisent pas d'analyse approfondie [33]. De plus, la validité des courbes d'attention comme explication a récemment été remise en question [18]. Il est donc nécessaire d'investiguer l'utilité du mécanisme d'attention comme outil pour la transparence.

3 JEUX DE DONNÉES ET MODÈLE

Nous comparons tout d'abord les performances d'un modèle d'apprentissage profond précédemment proposé [14] aux approches plus classiques construites avec des descripteurs construits à la main comme celles proposées par [25, 27]. Dans cette section, nous décrivons le nouveau jeu de données constitué, les descripteurs utilisés et le choix du modèle d'apprentisage profond (HireNet) ainsi que les légères modifications effectuées à celui-ci.

3.1 Jeu de données

Dans un souci de comparaison avec [14], nous avons décidé de sélectionner un type d'emploi spécifique : les postes de vente. Après filtrage basé sur des titres de poste spécifiques de la base de données ROME [1], une liste des postes a été sélectionnée et vérifiée par les auteurs et un expert des Ressources Humaines (RH). Enfin, en collaboration avec un acteur du secteur RH, nous avons obtenu un ensemble de données d'entretiens vidéo français comprenant plus de 627 postes. La retranscription automatique de la parole est obtenue grâce à l'API Google speech-to-text[2]. Il est important de noter que les vidéos sont très différentes de ce qui pourrait être produit dans une configuration de laboratoire. Les vidéos peuvent être enregistrées à partir d'une webcam, d'un smartphone ou d'une tablette, ce qui signifie que les environnements bruyants et les équipements de mauvaise qualité sont au rendez-vous. En raison de ces conditions réelles, l'extraction de fonctionnalités peut échouer pour une modalité pendant toute la réponse d'un candidat. Un exemple est la détection des unités d'action lorsque l'image a des problèmes d'éclairage. En ce sens, nous supprimons les vidéos si 1) la confiance retournée par la reconnaissance automatique de la parole est inférieure à 85 %, 2) OpenFace n'a pas réussi à détecter un visage dans plus de 20 % du total des fenêtres. Enfin, nous ne gardons que les candidats avec au moins une réponse pour laquelle le pipeline d'extraction des descripteurs a réussi pour les trois modalités. Certaines statistiques sur l'ensemble de données sont disponibles dans le tableau 1. Bien que les candidats aient accepté l'utilisation de leurs entretiens, l'ensemble de données ne sera pas rendu public en dehors du champ de cette étude car les vidéos sont des données personnelles soumises à de fortes contraintes de confidentialité.

Étiquetage des données et mesures d'évaluation. Les recruteurs regardent les vidéos des candidats et peuvent "liker" ou "disliker" les candidats, les présélectionner, les évaluer sur des critères prédéfinis ou écrire des commentaires. Pour simplifier la tâche, nous avons mis en place une classification binaire : les candidats présélectionnés, ayant reçu une opinion positive ou une note élevée sont considérés comme *Employable*, les autres candidats sont considérés comme *Non employable*. À noter ici que l'étiquette *Employable* traduit uniquement le souhait du recruteur d'inviter le candidat en entretien face-à-face. Si plusieurs annotateurs ont annoté le même candidat, nous procédons à un vote majoritaire. Nous avons divisé l'ensemble de données en un ensemble d'apprentissage, un ensemble de validation pour la sélection d'hyperparamètres et un ensemble de test pour l'évaluation finale de chaque modèle. Chaque ensemble constitue respectivement 70 %, 15 % et 15 % de l'ensemble de données complet.

Jeu de données	Train	Val	Test
Nombre de candidats	3581	784	783
Nombre moyen de questions	5,43	5,46	5,41
Nombre de postes	455	225	219
Proportion des étiquettes *Employable*	55 %	55 %	54 %

Table 1: Nombre de candidats et statistiques à propos de chacun des jeux de données.

3.2 Descripteurs utilisés

Nous avons trois sources d'information : les vidéos réponses dont nous extrayons les descripteurs textuels, prosodiques et faciaux, les titres des questions et le titre du poste. Pour chaque modalité d'une réponse vidéo, nous avons sélectionné des descripteurs de bas niveau. **Contenu verbal** : Nous avons choisi d'utiliser des plongements de mots contextualisés comme représentation du contenu verbal. De nombreuses tâches ont été améliorées en utilisant ces plongements de mots contextualisés notamment des taches d'inférence du langage naturel [21]. En tant que représentation pour les mots contextualisés, nous avons choisis d'utiliser une représentation produite par le modèle français pré-entraîné "CamemBERT" [21]. Nous obtenons ainsi comme représentation pour chacun des mots un vecteur de dimension 768. **Expression faciale** : Nous extrayons des descripteurs visuels pour chacune des fenêtres grâce à l'outil OpenFace [1], un logiciel d'analyse comportementale visuelle à la pointe de la technologie. Nous avons choisi d'extraire la position et la rotation de la tête, l'intensité et la présence des unités d'actions et la direction du regard. Comme de nombreuses vidéos ont des fréquences d'échantillonnage différentes, nous avons décidé de lisser les valeurs avec une fenêtre temporelle de 0,5s et un chevauchement de 0,25s. **indices prosodiques** : nos descripteurs prosodiques au niveau de la fenêtre sont extraits à l'aide de l'outil OpenSmile [10]. La configuration que nous utilisons est la même que celle utilisée pour obtenir les descripteurs eGeMAPS [9]. GeMAPS est un célèbre ensemble minimaliste de descripteurs sélectionnées pour leur importance dans l'informatique sociale, et eGeMAPS en est sa version étendue. Nous extrayons les descripteurs pour chaque fenêtre de 0,01s. Nous lissons ensuite ,de la même manière que pour la modalité vidéo, ces descripteurs prosodiques. **Questions et titre du poste**: Nous utilisons la représentation au niveau de la phrase du même modèle BERT préalablement utilisé pour représenter l'intitulé des questions et du titre de poste résultant en un vecteur de dimension 768.

[1] https://www.data.gouv.fr/en/datasets/repertoire-operationnel-des-metiers-et-des-emplois-rome/
[2] https://cloud.google.com/speech-to-text

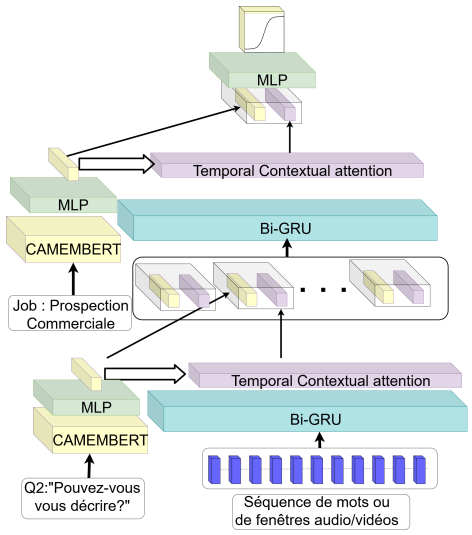

Figure 1: HireNet. Le bloc MLP correspond ici à un réseau de neurones à une couche. Le bloc Bi-GRU correspond à un réseau de neurones Bi-directionnel Gated Reccurent Unit.

3.3 HireNet

Dans un article précédent, nous avons proposé HireNet, un réseau de neurones d'attention pour prédire l'employabilité telle qu'évaluée par les recruteurs à partir d'entretiens vidéo structurés. HireNet [14] a été conçu pour représenter une séquence de questions-réponses contenant elles-mêmes une séquence de signaux sociaux ou de mots. Un schema rappelant l'architecture utilisé est disponible en figure 1. Nous avons procédé à plusieurs modifications vis-à-vis de l'architecture de base. Ainsi, nous avons 1) remplacé l'encodeur des questions et des titres de job par une simple couche encodant la représentation CamemBERT préentrainée; 2) modifié la fonction du mécanisme d'attention pour que le contexte soit mieux pris en compte. Pour rappel, le mécanisme d'attention bas niveau est destiné à discerner les moments les plus importants pour l'entretien d'embauche en fonction du contexte de l'intitulé de la question, tandis que le mécanisme haut niveau est destiné à pondérer les questions-réponses les plus importantes par rapport au contexte de l'intitulé du poste.

Ce nouveau mécanisme d'attention est décrit ci-dessous. Pour une séquence h_t et un vecteur de contexte fixe Q, l'attention est calculée de la façon suivante.

$$u_i = tanh(\lambda_i(W_q Q(W_h h_i)^\top) + (1-\lambda_i)(b^\top(W_k h_i))) \quad (1)$$

$$\lambda_i = \sigma([W_q Q * W_h h_i; b * W_k h_i] \quad (2)$$

$$\alpha_i = \frac{exp(u_i)}{\sum_{i'} exp(u_{i'})} \quad (3)$$

$$v = \sum_i \alpha_i h_i \quad (4)$$

où W_q, W_h, W_k sont des matrices de poids, b est un vecteur de poids, σ désigne la fonction sigmoïde, ; désigne la concatenation et $*$ désigne l'opération de produit terme à terme.

Baseline naïve	AUC			F1		
Aléatoire	0,5			0,5		
Vote majoritaire	0,5			0,354		
Vote par poste	0,630			0,627		
Modèle	Texte		Audio		Video	
	AUC	**F1**	*AUC*	**F1**	*AUC*	**F1**
Non séquentiel	0,613	0,570	0,624	0,584	0,579	0,549
HireNet	**0,727**	**0,659**	**0,738**	**0,662**	**0,661**	**0,612**

Table 2: Résultats sur la tâche de classification de l'employabilité

Le système de porte (λ) permet pour chaque instant de pondérer l'adéquation de cet instant avec le contexte par rapport à son importance intrinsèque dans le modèle. Par exemple, dans un contexte où une question liée à la relation client est posée, les mots en relation avec le concept de relation client devraient être mis en valeur (lambda élevé). Cependant il y a aussi des mots qui doivent être considérés comme important indépendamment du contexte, comme les insultes, les mots remplisseurs, l'usage du "je", etc. Dès lors le modèle peut apprendre l'importance des fenêtres ou des mots qu'il rencontre dépendamment de leur contexte ou de leur valeur intrinsèque respectivement grâce au premier et au second terme.

3.4 Évaluation expérimentale des classifieurs

Nous proposons ici une comparaison du modèle HireNet à plusieurs baselines : *i)* **vote aléatoire** (un millier de tirages aléatoires respectant l'équilibre des étiquettes du jeu de données d'entraînement ont été effectués. Le score AUC est ensuite moyenné ; *ii)* **Vote par majorité** (cette méthode consiste simplement à attribuer l'étiquette majoritaire ; *iii)* **Vote par poste** (cette méthode consiste simplement à attribuer l'étiquette majoritaire du poste concerné. Puisque notre modèle pourrait simplement apprendre l'étiquette la plus représentée pour chacun des postes, nous avons décidé d'inclure ce modèle ; *iv)* **Méthodes classiques non séquentielles** nous appliquons des fonctions statistiques pour réduire la dimension temporelle et pour obtenir un vecteur fixe pour chaque modalité. Les fonctionnelles moyenne, écart type, minimum, maximum, somme des gradients positifs et somme des gradients négatifs sont appliqués aux descripteurs décrits dans la partie 3.2 pour obtenir une représentation de chaque réponse pour l'audio et la vidéo. Pour la modalité de langage nous prenons la moyenne des représentations CamemBERT. Trois algorithmes d'apprentissage classiques (à savoir SVM, régression Ridge et forêt aléatoire) sont utilisés pour la baseline. Le meilleur résultat des trois algorithmes est retenu. Les métriques d'évaluation choisies sont l'aire sous la courbe (AUC) et la moyenne du score F1 des classes *Employable* et *Non employable*. Les résultats sont reportés dans le tableau 2. Nous constatons que HireNet présente des meilleurs résultats par rapport aux méthodes non séquentielles. De plus, il est intéressant de noter que les modalités du langage et de la prosodie sont les modalités qui obtiennent les meilleurs résultats quel que soit le classifieur utilisée.

Modalité	Audio			Texte			Vidéo		
attention slices détectées	62400			65056			45636		
attention slices par question (moyenne)	3,20			2,61			2,23		
attention slices par entretien (moyenne)	12,63			12,86			9,54		
	Q10%	Moyenne	Q90%	Q10%	Moyenne	Q90%	Q10%	Moyenne	Q90%
Durée des *attention slices*	0,5s	1,2s	2s	2 mots	5,8 mots	10 mots	0,5s	1,45s	2,5s

Table 3: Statistiques descriptive des *attention slices* détectées. Q10% et Q90% dénotent les quantiles à 10 et 90 %

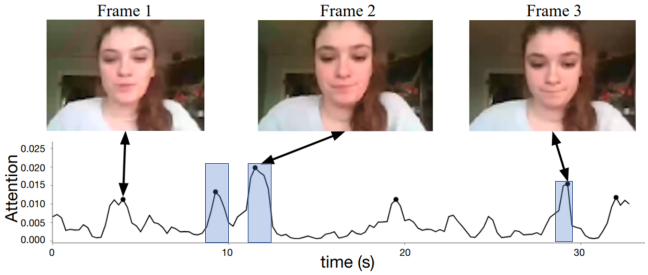

Figure 2: Exemple d'une courbe d'attention et de moments saillants détectés par HireNet.

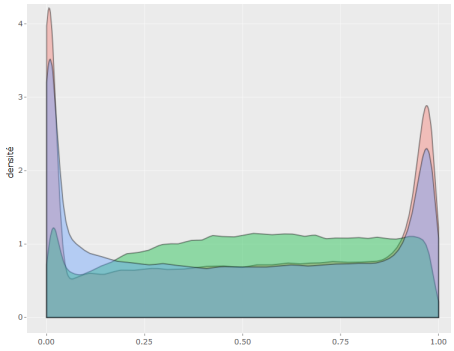

Figure 3: Densité des *attention slices* en fonction de leur moment d'apparition au regard de la longueur totale de réponse. Les courbes bleu, rouge et verte représentent respectivement les modalités vidéo, audio et texte

4 ANALYSE DES *ATTENTION SLICES*

4.1 Les courbes d'attention exposent-elles réellement des pics distincts?

Comme les réseaux de neurones sont soumis à diverses sources de variabilité telles que l'initialisation des poids aléatoires, la descente du gradient stochastique ou l'utilisation des méthodes de *dropout*, nous avons choisi d'entraîner cinq instances du modèle, puis de calculer la moyenne des valeurs d'attention. La moyenne des valeurs de ces courbes est ensuite utilisée pour le reste de l'étude. De cette façon, nous visons à capturer le comportement plus général des mécanismes d'attention [19]. Pour le reste de l'article, nous définissons une *attention slice* comme une tranche sélectionnée en fonction de la courbe d'attention.

Méthodologie : Extraction des attention slices par détection des valeurs aberrantes de façon non supervisée. Les courbes d'attention consistent principalement en des séries temporelles bruitées avec des pics de valeur élevée [33]. Un exemple typique de courbes d'attention est affiché en la figure 2. La première étape de notre méthodologie consiste à filtrer les moments où ces pics d'attentions augmentent (*attention slices*). Pour ce faire, nous utilisons l'algorithme DBSCAN [8], un algorithme de clustering par densité qui nous permet de sélectionner les régions d'attention maximale, illustrées par les boîtes bleues sur la figure 2.

Résultats et statistiques descriptives sur les pics extraits. Le tableau 3 fournit un résumé des données servant de base à notre étude en termes de pics d'attention détectés et leurs statistiques descriptives. Tout d'abord, il est intéressant de noter que la durée des tranches importantes extraites par le mécanisme de l'attention pour les modalités audio et vidéo sont principalement comprises entre 0,5s et 2,5s, durée qui parait en adéquation avec la durée d'une expression faciale ou d'une disfluence. À noter que les *attention slices* audio semblent être plus courtes que celles des expressions faciales. Concernant le texte, des extraits de moins d'une dizaine de mots sont extraits. De plus, moins d'*attention slices* sont détectées pour la modalité vidéo que pour les modalités audio et textuelle. Ensuite, nous fournissons les courbes de densité pour les instants d'apparitions des *attention slices* au regard de la durée totale de la réponse 3. Les *attention slices* pour les modalités audio et vidéo se produisent plus souvent au début et à la fin d'une réponse. De tels indices pourraient indiquer que les comportement non verbaux survenant au début (prise de parole) et à la fin de la réponse ont un impact important sur la décision d'employabilité, comme dans les interactions face-à-face [3, 13].

4.2 Les signaux sociaux lors des *attention slices* sont-ils différents de ceux des tranches aléatoires ?

Méthode: Classification supervisée entre les attention slices et des tranches aléatoires. Les pics d'attention peuvent fortement dépendre du contexte et de la mémoire des cellules GRU et très peu des descripteurs d'entrée au moment de leur apparition. Nous émettons l'hypothèse qu'un pic d'attention est dû à un changement de comportement dans une réponse et donc dépend principalement du pas du temps traité. Pour nous assurer que les pics d'attention proviennent principalement de ce qui se passe dans les pas de temps concernés, nous construisons une tâche de classification binaire. Cette tâche consiste à distinguer les moments à fortes attentions d' autres moments aléatoires en se basant sur les comportements verbaux et non verbaux se produisant durant ceux-ci. Ainsi, pour notre tâche, nous prenons comme étiquette positive les *attention*

	Texte			Audio			Video		
Modèle	F1 positif	F1 négatif	**AUC**	F1 positif	F1 négatif	**AUC**	F1 positif	F1 négatif	**AUC**
Aléatoire	0,199	0,799	0,499	0,199	0,799	0,499	0,199	0,799	0,499
Vote majoritaire	0	0,888	0,5	0	0,888	0,5	0	0,888	0,5
Ridge	0,3535	0,6142	0,7203	0,7798	0,9456	0,9615	0,5589	0,9131	0,8738
XGBoost	0,1749	0,8887	**0,7553**	0,8297	0,9550	**0,9725**	0,6203	0,9177	**0,9069**

Table 4: Résultats de classification entre les *attention slices* et les tranches aléatoires

slices extraites dans chacune des réponses du candidat. Comme étiquette négative, pour chacune des *attention slices* détectées, nous échantillons quatre moments dans la réponse du candidat avec la même durée que l' *attention slice* précédemment sélectionnée. Notre objectif étant de comprendre si les *attention slices* sont différentes et en quoi elles diffèrent, nous avons décidé d'utiliser les classifieurs standards Ridge et XGBoost [5], pour lesquels une méthodologie d'analyse d'importance des descripteurs est bien établie. Comme ces classifieurs prennent comme entrée un vecteur fixe, nous construisons pour chacune des modalités un set de descripteurs interprétables. **Pour les descripteurs audio et vidéos**, nous intégrons les descripteurs des fenêtres temporelles des moments choisis en utilisant les fonctionnelles suivantes: moyenne, moyenne des gradients positifs et moyenne des gradients négatifs. Ces fonctionnelles sont appliquées sur le même ensemble de descripteurs que celui utilisé dans la sous-section 3.2 pour entraîner notre précédent modèle. L'unique différence réside dans l'application d'une étape de pré-traitement de Z-normalisation par rapport à la réponse complète. **Pour les descripteurs textuels**, nous construisons un set de descripteurs compréhensibles pouvant nous informer sur quels aspects, une *attention slice* dérive du contenu verbal. Pour cela, nous utilisons des méthodes à base de dictionnaires notamment, le LIWC qui classent des mots selon des catégories haut niveau (ie travail, accomplissement, etc), le dictionnaire FEEL qui catégorise des mots dans l'une des 6 emotions de base définies par Ekman, le dictionnaire LEXIQUE3 qui fournit des informations quant à la fréquence de l'utlisation des mots dans la langue francaise et la nature grammaticale des mots utilisés obtenu grâce à l'outil TreeTagger. Ces dictionnaires ont déjà prouvé leur utilité précédemment dans différentes tâches liées à la prédiction de l'employabilité [15]. Pour cette expérience, nous conservons les mêmes ensembles d'entraînement, de développement et de tests que dans la section 3.1 pour éviter toute fuite de données.

Résultats et analyses. Comme indiqué dans le tableau 4, les performances des classifieurs choisis présentent debons résultats pour les modalités audio et vidéo, ce qui prouve que, malgré l'influence de la modélisation de séquence et de l'utilisation des informations de contexte, l'importance d'un moment est encore principalement définie par les événements qui s'y produisent. Cela montre que les moments précis où se produisent des pics d'attention se distinguent des autres moments de la même réponse. Concernant la modalité textuelle, les résultats sont plus mitigés. L'AUC ne dépasse pas le score de 0.75 et, bien que le résultat soit meilleur que les modèles aléatoires et vote majoritaire, nous pouvons émettre des réserves quant à la possibilité de détecter les moments importants en utilisant notre set de descripteurs. En effet, aucun descripteur sémantique n'est utilisé et la représentation vectorielle utilisée n'est peut être pas assez riche comparée à celle par plongement de mot.

4.3 Analyse de l'importance des descripteurs discriminant les *attention slices*

Une expérience sur l'importance des descripteurs a également été réalisée afin de mettre en évidence les descripteurs qui contribuent le plus lors de l'identification des *attention slices.* Ainsi, nous inspectons les coefficients du modèle Ridge et nous les classons en fonction de leur magnitude. En ce qui concerne l'importance pour le modèle de XGBoost, nous inspectons les descripteurs ayant contribué le plus à diminuer l'impureté à chaque embranchement. Le tableau 5 présente, pour chacune des modalités, les dix descripteurs les plus discriminants pour les deux méthodes. **Pour les expressions faciales**, les deux principaux descripteurs du bas du visage mis en exergue par notre méthode sont respectivement l'activation de l'AU17 (l'élévation du menton) et la non activation de l'AU26(l'ouverture de la mâchoire). Les mouvements liés au sourcil ont aussi été détectés comme discriminants notamment les moments où l'AU02 (Remontée de la partie externe des sourcils) est moins activée. Enfin les rotations de la tête et plus précisément le tangage (Rx), et le roulis (Rz) sont détectés comme des descripteurs importants. **Pour les indices audios**, deux grands groupes de descripteurs émergent notamment ceux associés à la prosodie et les descripteurs de la qualité de la voix et spectraux. Le premier groupe semble indiquer que les silences sont souvent considérés comme des moments importants(coefficient négatif pour loudness), tandis que le deuxième groupe semble lié à la respiration et à la tension dans la voix (harmonic noise ratio, logrelf0.h1.a3). **Pour les indices de langage**, aucun descripteur n'a une valeur d de cohen supérieure à 0.5. Cependant, il existe un faible effet pour la singularité lexicale (utilisation de mots plus ou moins fréquent dans la langue francaise), les pronoms et pronoms personnels.

4.4 Les *attention slices* portent-elles plus d'informations par rapport à la tache de prédiction d'employabilité que des tranches aléatoires?

Méthode : Classification supervisée concernant l'employabilité grâce aux tranches fines aléatoires ou aux attention slices. Nous avons établi que notre modèle pouvait mettre en exergue un certain nombre de moments dans la réponse du candidat et avons constaté que les *attention slices* étaient différentes des autres tranches de la réponse. Nous étudions maintenant leur utilité pour la tâche de prédiction de l'employabilité, en vérifiant que les moments contenus dans ces *attention slices* ont un contenu prédictif supérieur par rapport à des

Groupe	**Ridge**		**XGBoost**
	Coefficients positifs	Coefficients négatifs	
Bas du visage	**AU17**[3], AU14[9]	AU10[2], **AU26**[7], AU20[8], AU28[10],	**AU17**[6],**AU26**[7],AU10[8],**AU25**[10]
Haut du visage	**AU02**[5]	AU04[4]	**AU02**[1],AU04[9]
Position et rotation de la tête	**R_z**[1], **R_x**[6],		**R_x**[3], **R_z**[4]
Confidence d'OpenFace			↑success[2], success[5]
Descripteurs cepstraux		**mfcc2**[2], **mfcc3**[8]	**mfcc2**[5],↑mfcc2[10]
Descripteurs spectraux	f1bandwidth[9]	**spectralflux**[1], **f1amplitude**[7], f3frequency[10]	**f1amplitude**[2], **spectralflux**[4], **f3amplitude**[7], ↑**spectralflux**[8], f1bandwidth[9]
Descripteurs de la prosodie		**loudness**[5]	**loudness**[6]
Descripteurs de la qualité de la voix	**h1.a3**[4], **slope500.1500**[5], **h1.h2**[6]	**hnr**[3]	**hnr**[1], **slope500.1500**[3]
LIWC			je[2], verbeprésent[5], verbefutur[10]
LEXIQUE3	freqlemfilms_sd[2], freqlemlivres_mean[3], freqlemlivres_max[10]	freqlemfilms_mean[1], freqlemlivres_sd[4], freqlemfilms_max[8]	freqlemfilms_max[3], nbsyll_max[4], freqlemfilms_q25[6], nbhomoph_max[9]
Nature Grammaticale		aux[5], verb[6], pron[7], det[9]	pron[1], aux[7], sconj[8]

Table 5: Analyse de l'importance des descripteurs

F^i indique que le descripteur F est classé en i_{eme} position. ↑ et ↓ représentent respectivement la moyenne des gradients positifs et négatifs. Les descripteurs en gras **F** ont une taille d'effet au minimum "moyenne" basée sur un test d de cohen $d > 0.5$

	Texte		Audio		Video	
	Tranche aléatoire	*attention slice*	Tranche aléatoire	*attention slice*	Tranche aléatoire	*attention slice*
Ridge	0,5259 ± 0,0171	0,5296 ± 0,0190	0,5092 ± 0,0246	0,5445 ± 0,0158	0,5149 ± 0,0182	0,5163 ± 0,0308
XGBoost	0,5179 ± 0,0199	0,5245 ± 0,0233	0,5142 ± 0,0227	0,5650 ± 0,0235	0,5151 ± 0,0192	0,5185 ± 0,0220

Table 6: Résultat de la tâche de classification de l'employabilité à partir des *attention slices* ou de tranches aléatoires

tranches aléatoires. En utilisant les mêmes descripteurs que dans la sous-section 4.2, nous construisons une tâche de classification basée sur une fenêtre temporelle minimale dans la réponse du candidat. Pour ce faire, nous constituons deux sous-ensembles du jeu de données précédent : la première ne contient que les *attention slices* ; la seconde est composée des tranches aléatoires sélectionnées en section 4.2. Pour chacun des sous-ensembles du jeu de données, nous effectuons la procédure de *bootstrapping* suivante : nous formons 100 nouvelles instances d'ensemble d'apprentissage, chacune étant un échantillonage composé d'une seule *attention slice* ou tranche aléatoire par candidat respectivement pour le premier et le deuxième sous-ensemble. Nous entraînons les classifieurs Ridge et XGBoost sur ces jeu de données et les testons sur deux sous-ensembles d'une seule *attention slice* ou tranche aléatoire par candidat respectivement pour le premier et le deuxième sous-ensemble. Ainsi, nous obtenons un ensemble de scores d'AUC permettant de calculer l'intervalle de confiance des moyennes de nos résultats pour avoir une idée de leur significativité statistique. A noter, que nous respectons toujours les mêmes divisions d'entraînement, de développement et de tests que dans la section 3.1 pour éviter toute fuite de données.

Résultats et discussions. Les résultats sont rapportés dans le tableau 6. Nous observons une moyenne d'AUC significativement plus élevée (*p-values* < 0.01) pour les *attention slices* que pour les *tranches* aléatoires concernant la modalité audio, et ce pour les deux classifieurs. En revanche, les modalités texte et vidéo ne montre pas de différence significative de performances. Nous concluons que les *attention slices* contiennent effectivement plus d'information permettant de prédire l'employabilité du candidat pour la modalité audio.

5 CONCLUSION

Dans cet article, nous étudions la possibilité d'utilisation d'un modèle d'apprentissage profond HireNet pour la prédiction de l'employabilité de candidats dans le cadre d'entretiens d'embauche vidéo différés. Nous utilisons séparément trois modalités (langage, prosodie et expressions faciales) et nous comparons ce modèle à des approches de la littérature sur un nouveau jeu de données réelles. Nos expériences montrent que HireNet obtient des résultats significativement meilleurs que les autres approches. Cependant, un tel modèle troque ce gain de performances contre une opacité plus importante. Pour pallier cette opacité, nous proposons de nous appuyer sur le mécanisme d'attention intrinsèque au modèle pour expliquer les prédictions. Pour ce faire, nous extrayons les tranches contenant des pics d'attention, qui correspondent aux périodes de l'entretien jugées les plus significatives par le modèle. Nous montrons d'abord que ces tranches apparaissent le plus souvent au début et à la fin des réponses pour les modalités audio et vidéo. D'autre part, nous avons qualifié ces moments en termes de comportements verbaux et non verbaux grâce à une étude d'importance des descripteurs. Nous montrons enfin que les tranches correspondant à la modalité audio conservent une faible valeur prédictive en dépit de leur courte

durée. Nous n'observons pas ce résultat pour les deux autres modalités, ce qui pourrait être dû à une information insuffisante dans une unique tranche. Les valeurs d'attention mettent en évidence l'importance des moments, mais n'incluent pas d'informations sur s'ils ont un impact positif ou négatif sur les décisions des recruteurs, ce qui pourra faire l'objet de travaux futurs. De plus, comme notre approche est basée sur un modèle appris (HireNet), une direction de recherche consiste à l'améliorer en termes de performances et de contrôle des possibles biais. Enfin, nous prévoyons de mener une tâche d'annotation et une étude utilisateur afin de: *i)* quantifier la transparence et l'utilité perçues d'HireNet et des *attention slices* extraites ; *ii)* créer une plateforme automatique capable de fournir des commentaires utiles aux candidats pour l'entraînement à l'entretien d'embauche.

6 REMERCIEMENTS

Ce travail a été supporté par la société EASYRECRUE, nous tenions à remercier Florian Chaux et Amandine Reitz pour leur soutien.

REFERENCES

[1] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis Philippe Morency. 2018. OpenFace 2.0: Facial behavior analysis toolkit. *Proceedings - 13th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2018* (2018), 59–66. https://doi.org/10.1109/FG.2018.00019

[2] Dana R. Carney, C. Randall Colvin, and Judith A. Hall. 2007. A thin slice perspective on the accuracy of first impressions. *Journal of Research in Personality* 41, 5 (10 2007), 1054–1072. https://doi.org/10.1016/j.jrp.2007.01.004

[3] Justine Cassell, Yukiko I Nakano, Timothy W Bickmore, Candace L Sidner, and Charles Rich. 2001. Non-verbal cues for discourse structure. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics - ACL '01*. Association for Computational Linguistics, Morristown, NJ, USA, 114–123. https://doi.org/10.3115/1073012.1073028

[4] Lei Chen, Ru Zhao, Chee Wee Leong, Blair Lehman, Gary Feng, and Mohammed Ehsan Hoque. 2017. Automated video interview judgment on a large-sized corpus collected online. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 504–509.

[5] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 13-17-Augu (2016), 785–794. https://doi.org/10.1145/2939672.2939785

[6] Mathieu Chollet and Stefan Scherer. 2017. Assessing Public Speaking Ability from Thin Slices of Behavior. *Proceedings - 12th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2017 - 1st International Workshop on Adaptive Shot Learning for Gesture Understanding and Production, ASL4GUP 2017, Biometrics in the Wild, Bwild 2017, Heteroge* (2017), 310–316.

[7] Timothy Degroot and Janaki Gooty. 2009. Can nonverbal cues be used to make meaningful personality attributions in employment interviews? *Journal of Business and Psychology* 24, 2 (2009), 179–192.

[8] Martin Ester, Hans-peter Kriegel, Xiaowei Xu, and D Miinchen. 1996. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. (1996).

[9] Florian Eyben, Klaus R. Scherer, Bjorn W. Schuller, Johan Sundberg, Elisabeth Andre, Carlos Busso, Laurence Y. Devillers, Julien Epps, Petri Laukka, Shrikanth S. Narayanan, and Khiet P. Truong. 2016. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Transactions on Affective Computing* 7, 2 (2016), 190–202.

[10] Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller. 2013. Recent developments in openSMILE, the munich open-source multimedia feature extractor. *Proceedings of the 21st ACM international conference on Multimedia - MM '13* May (2013), 835–838. https://doi.org/10.1145/2502081.2502224

[11] Amanda R. Feiler and Deborah M. Powell. 2016. Behavioral Expression of Job Interview Anxiety. *Journal of Business and Psychology* 31, 1 (2016), 155–171. https://doi.org/10.1007/s10869-015-9403-z

[12] Ray J. Forbes and Paul R. Jackson. 1980. Non-verbal behaviour and the outcome of selection interviews. *Journal of Occupational Psychology* 53, 1 (1980), 65–72. https://doi.org/10.1111/j.2044-8325.1980.tb00007.x

[13] Wells Goodrich. 1979. Face-to-Face Interaction: Research, Methods, and Theory. *Family Process* 18, 3 (9 1979), 355–356.

[14] Leo Hemamou, Ghazi Felhi, Vincent Vandenbussche, Jean-claude Martin, and Chloe Clavel. 2019. HireNet : a Hierarchical Attention Model for the Automatic Analysis of Asynchronous Video Job Interviews. In *AAAI*.

[15] Léo Hemamou, Grégory Wajntrob, Jean-claude Martin, and Chloé Clavel. 2018. Entretien vidéo différé : modèle prédictif pour pré-sélection de candidats sur la base du contenu verbal. In *Workshop sur les "Affects, Compagnons Artificiels et Interactions" (ACAI)*. 1–8.

[16] Mohammed Ehsan Hoque, Matthieu Courgeon, Jean-Claude Martin, Bilge Mutlu, and Rosalind W. Picard. 2013. MACH. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing - UbiComp '13*. ACM Press, New York, New York, USA, 697. https://doi.org/10.1145/2493432.2493502

[17] Allen I Huffcutt, James M Conway, Philip L Roth, and Nancy J Stone. 2001. Identification and meta-analytic assessment of psychological constructs measured in employment interviews. *Journal of Applied Psychology* 86, 5 (2001), 897–913. https://doi.org/10.1037//0021-9010.86.5.897

[18] Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. In *North American Chapter of the Association for Computational Linguistics*. http://arxiv.org/abs/1902.10186

[19] Saumya Jetley, Nicholas A Lord, Namhoon Lee, and Philip H S Torr. 2018. Learn To Pay Attention. In *International Conference on Learning Representations*. 1–14. http://arxiv.org/abs/1804.02391

[20] Xuan Liu, Xiaoguang Wang, and Stan Matwin. 2019. Improving the interpretability of deep neural networks with knowledge distillation. *IEEE International Conference on Data Mining Workshops, ICDMW* 2018-Novem (2019), 905–912. https://doi.org/10.1109/ICDMW.2018.00132

[21] Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah, and Benoît Sagot. 2019. CamemBERT: a Tasty French Language Model. 2 (2019). http://arxiv.org/abs/1911.03894

[22] Skanda Muralidhar, Laurent Nguyen, and Daniel Gatica-Perez. 2019. Words Worth: Verbal Content and Hirability Impressions in YouTube Video Resumes. (2019), 322–327. https://doi.org/10.18653/v1/w18-6247

[23] Nora A. Murphy, Judith A. Hall, Marianne Schmid Mast, Mollie A. Ruben, Denise Frauendorfer, Danielle Blanch-Hartigan, Debra L. Roter, and Laurent Nguyen. 2015. Reliability and Validity of Nonverbal Thin Slices in Social Interactions. *Personality and Social Psychology Bulletin* 41, 2 (2 2015), 199–213. https://doi.org/10.1177/0146167214559902

[24] Iftekhar Naim, Md. Iftekhar Tanveer, Daniel Gildea, and Mohammed Ehsan Hoque. 2018. Automated Analysis and Prediction of Job Interview Performance. *IEEE Transactions on Affective Computing* 9, 2 (4 2018), 191–204. https://doi.org/10.1109/TAFFC.2016.2614299

[25] Laurent Son Nguyen and Daniel Gatica-Perez. 2015. I Would Hire You in a Minute. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction - ICMI '15*. ACM Press, New York, New York, USA, 51–58. https://doi.org/10.1145/2818346.2820760

[26] Pooja Rao S. B, Sowmya Rasipuram, Rahul Das, and Dinesh Babu Jayagopi. 2017. Automatic assessment of communication skill in non-conventional interview settings: a comparative study. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction - ICMI 2017*. ACM Press, New York, New York, USA, 221–229. https://doi.org/10.1145/3136755.3136756

[27] Sowmya Rasipuram and Dinesh Babu Jayagopi. 2018. Automatic assessment of communication skill in interview-based interactions. *Multimedia Tools and Applications* 77, 14 (2018), 18709–18739. https://doi.org/10.1007/s11042-018-5654-9

[28] Marco Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*. Association for Computational Linguistics, Stroudsburg, PA, USA, 97–101. https://doi.org/10.18653/v1/N16-3020

[29] Neal Schmitt (Ed.). 2012. *The Oxford Handbook of Personnel Assessment and Selection*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199732579.001.0001

[30] Leann Schneider, Deborah M. Powell, and Nicolas Roulin. 2015. Cues to deception in the employment interview. *International Journal of Selection and Assessment* 23, 2 (2015), 182–190. https://doi.org/10.1111/ijsa.12106

[31] George Trigeorgis, Fabien Ringeval, Raymond Brueckner, Erik Marchi, Mihalis A. Nicolaou, Bjorn Schuller, and Stefanos Zafeiriou. 2016. Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings* 2016-May (2016), 5200–5204. https://doi.org/10.1109/ICASSP.2016.7472669

[32] Fan Yang, Mengnan Du, and Xia Hu. 2019. Evaluating Explanation Without Ground Truth in Interpretable Machine Learning. (2019). http://arxiv.org/abs/1907.06831

[33] Hongliang Yu, Liangke Gui, Michael Madaio, Amy Ogan, Justine Cassell, and Louis-Philippe Morency. 2017. Temporally Selective Attention Model for Social and Affective State Recognition in Multimedia Content. *Proceedings of the 2017 ACM on Multimedia Conference* (2017), 1743–1751.