ML4Science: Psychology takes on ML

Adam Narozniak, Noé Durandard and Hugo Casademont
CS-433 Machine Learning, EPFL, Switzerland

***Abstract*—The adoption of asynchronous video interviews (AVIs) in job interviews ultimately involves machines in employment processes. The role and powers of computational methods to analyse these negotiations, often sprinkled with deception, need to be transparently studied and assessed. This paper explores the capacity of low- and high-level features extracted from interviews' transcripts and audio recordings to align with human-rated performance and deception levels.**

I. Introduction

The recent emergence and spread of asynchronous video interviews (AVIs) asks for the transparent examination of machine learning methods to assess candidates' performances. HireVue, the market leader in this sector, states that more than one-third of Fortune 100 companies use their services, and this figure is growing [1]. Every day thousands of applicants get evaluated by their methods. Considering these circumstances, it is crucial to better understand and evaluate how machines can encompass and chart traditional human-analysed cues during job interviews. This project, led in collaboration with Industrial/Organizational (I/O) psychologists Dr Bangerter, Professor at University of Neuchâtel (Switzerland), and Dr Roulin Professor at Saint Mary's University (Halifax, Canada), aims at investigating this issue. This paper presents the computational investigation of 800 filmed and transcribed answers from simulated in-person job interviews. The extraction of multi-sources highly interpretable features from the transcripts and audio recordings are automated and could potentially be integrated into larger pipelines. Machine learning models are then trained to investigate the extracted features' effects on high-quality human ratings, putting the emphasis on applicants' overall performance and deception strategies.

II. Interview Datasets

The dataset studied in this project was kindly provided by Dr Roulin. It is composed of video recordings and transcripts of 100 applicants' answers to the same eight questions during experimental job interview simulations. The interviews are conducted in English. Interviewees were balanced between undergraduate students in Bachelor of Commerce (47%) and MBA graduate students (53%). This is included in the dataset as a binary class named Education Level. They have beforehand ascribed an IM condition for each of the questions as an instruction to give honest responses (35%) or to deceive slightly (35%) to extensively (30%) the interviewer.

The content of the interviews is enriched with several scores given to the answers by the interviewer or independent raters, such as interview performance or Criterion-Based Content Analysis (CBCA). The performance of the interviewees for each of the questions was assessed by the interviewer on a scale ranging from 1 to 5. CBCA was performed for each of the answers by two distinct raters. This is a composite measure monitoring the presence or absence of 14 different indicators (mainly based on the structure of the production, its amount of details, and spontaneity) [2]. Each indicator is coded as 0 if it is absent from the response, 0.5 for some indication of presence in the response and 1 if the response indicates the clear presence of the indicator. This results in a CBCA score ranging from 0 to 14 with 0.5-point increments. This indicator is widely used to "assess the veracity of written statements and is used as evidence in criminal courts in several countries in the world" [3].

III. Features extraction

We extracted a range of different features: lexical features from the transcripts, prosodic and emotion features from the audio and finally, text embedding of the transcripts using BERT [4]. Due to the time constraint and limited resources, we were, however, unable to extract features from the video frames. We based part of our extraction process on previous work where interviews were also analysed [5].

A. Lexical features

Various lexical features were extracted automatically from the transcripts of the interviews in order to grasp information regarding their verbal content, the topics developed and the applicants' speech style. Based on traditional computerised psycho-linguistic studies, Empath, an open-source alternative to the widely used LIWC tool [6] showcasing comparable results [7], was used to count words in psychologically meaningful categories, including emotions and many different semantic fields. Entities mentions (dates, places, organisations, etc.) were added to these statistics thanks to spaCy library [8] Manually extracted disfluency markers were also counted. Additional syntactic features were retrieved using ConvoKit's Politeness API, a toolkit to study social phenomena in conversations developed by Cornell University's NLP team [9], [10]. It allowed the extraction of several strategies such as the use of deference, gratitude, factuality, hedges, apologies, counterfactual and indicative modes. Then, in order to embrace the nature and specificities of the dataset at hand, topics were automatically extracted from the corpus, a priori, through gensim's implementation of Latent Dirichlet Allocation (LDA) [11]. The optimal number of topics was determined by minimising the U_{mass} coherence measure designed to quantify the coherence of topics [12], within three to thirty topics.

Finally, considering that the extracted features were, by construction, highly correlated to redundant and large compared to the number of instances, dimensionality reduction schemes were applied to the feature vectors to analyse their impact on fitted models. Uninformative features with relation to pre-determined target variables were removed based on Pearson correlation's p-value. Then two separated unsupervised feature space reductions were implemented. Principal Component Analysis (PCA) was performed while fitting the models. Features were also clustered a priori in a greedy manner through an agglomerative paradigm, grouping features based on their similarity. The optimal number of clusters was obtained by maximising the silhouette score [13], and the central feature of each cluster was retained for further analyses.

B. Prosodic features

Several prosodic features were extracted from the audio of the videos. The extraction process was based on the previous work presented above [5]. In the cited paper, the free computer software package for speech analysis in phonetics Praat was used [14]. We decided to use the python library parselmouth [15] which implements praat methods directly into python and allows automated extraction. In total, 39 features were extracted, including the mean, standard deviation of pitch, formant and intensity, as well as other measures such as pauses duration and breaks (see all features in Table II in the appendix).

C. Audio based emotions prediction

Assessment of interviewee emotions was shown to help predict job interview outcomes [5]. In the cited paper, features were obtained by Mechanical Turk workers' ratings. We decided to extract similar features by audio analysis. The emotions that we predict are neutral, happy, sad, angry, fearful, surprise, and disgust. Details of this process are described onward.

1) Data: In order to predict emotions, we gathered two datasets to train the model: RAVDESS [16] and TESS [17] (described in appendix subsection B).

2) Preprocessing: Preprocessing started with removing silent parts of the audio. Then based on the length of the audio clips, a standard length was established so that there is no need to clip the audio extensively (in order not to lose information) and not to introduce too much zero padding. Then audio clips were resampled to 16 kHz. Then denoising was applied. The following features were extracted from the audio: the Mel frequency cepstral coefficients, zero-crossing rate, and root mean square (energy).

3) model training: An artificial neural network model was designed to generate predictions based on Long-Short Term Memory (see architecture in Figure 1). The model achieved 88% accuracy. The mistakes are display using the confusion matrix in subsection C, Figure 3.

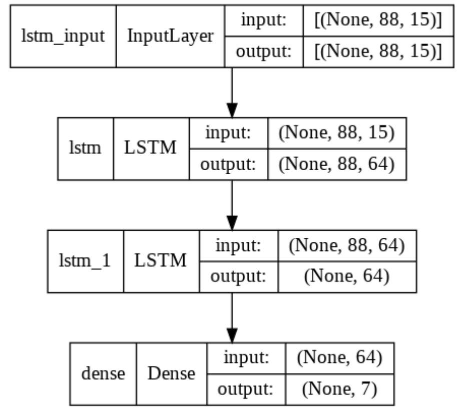


Figure 1: Emotions predictions model architecture.

4) results on the interview dataset: The interviews were preprocessed using the same procedures as while training. Then, the emotions were predicted for every clip of the length equal to the one used during the training. The probability of emotions belonging to a certain class was kept instead of the hard classification. Then in order to predict the interview performance, the emotions obtained for every point in time were averaged. More sophisticated methods can be further examined. Also standard deviation can be worth examining.

IV. Models and methods

A. Target variables

In the context of this project, we decided to focus on four target variables included in the original dataset (described in section II). Two of them, the Education Level and the IM condition, defined upstream from the interview, are treated as categorical variables. The two others, Interview performance and CBCA Score, result from the applicants' answers. Thus, the models inquire about the possibilities to understand both interviewees' intrinsic conditions and results based only on their production.

B. Models

After a coarse examination of machine learning models based on their potential interpretability or predictive power, we selected empirically three models from the scikit-learn library [18] to perform classification and their corresponding versions for regression. We also created models using the Tensorflow library [19] in order to fine-tune BERT model [4]. For each of the four preselected target variables, models were trained using different sets of independent variables (lexical, prosodic, predicted from audio emotions, and combinations of them) with these three models, namely ridge regression model [20], and two ensemble learning methods: random forests [21] and ADABoost [22]. Crucial parameters of these models were tuned through grid search paradigm[1] and evaluated using 5-fold cross-validation in case of the sklearn models and test set in case of Tensorflow models:

- Ridge: linear least squares regressor with a l^2-regularizer with strength controlled by the hyperparameter α optimized over 200 values ranging from 50 to 1000. Its classifier alternative converts targets variables to binary labels in a one-vs-all fashion for multi-class classification and trains a regressor for each label. In a multi-class scenario, the class is predicted to achieve the highest score.
- Random forest: meta-estimator that operates by fitting a multitude of decision trees on various sub-samples of the data and then averaging their results. The number of trees in the forest is optimised over 20 values from 50 to 250.
- ADABoost: meta-estimator fitting first a model on the original dataset (here with a decision tree) and then subsequent copies of the base model *adaptated* (weighted) to focus on instances mispredicted by the previous model. The learning rate of the algorithm is tuned over 50 values from 10^{-3} to 10^4.
- BERT-based model: it is a transformer-based model that uses a pre-trained version of BERT (with 4 hidden layers of size 512 and 4 attention heads) with one Dense Layer with one neuron for regression and n-1 neurons for classification. The data was divided into 70, 15, and 15 % for training, validation, and testing. The input length was truncated or zero-padded (depending on the case) to 512 tokens which is the maximum number of the tokens for pre-trained BERT models.

C. *Quality assessment*

During the grid search, two different metrics were optimised depending on the nature of the task. For classification, the accuracy of the model was maximised, while for regression, the Mean Absolute Error (MAE) was minimised. Tensorflow based models were optimised for MSE though the final metric remains the MAE for interpretability reasons. The average of these scores among the cross-validated folds was retained as the performance of the methods to account for the targeted phenomenon. In order to put these values into perspective, they have been benchmarked against dummy models: the average for regression tasks and the most frequent class in classification tasks.

[1]Due to limited computational power, the grid search was constrained to a limited number of parameters.

V. Results

The results obtained for the three models fitted with the different sets of features are gathered in Table I, for each target variable the best score is embolden and scores worse or equal to the one achieved with the dummy model are italicized. As the altered lexical features often attained worst results than their complete version, the full version was used when combined with other set of features, results are visible in subsection D, Table III.

The developed models have proved to be powerful to determine applicants' education level, reaching more than 20 points above chance using both prosodic and lexical features. Moreover, prosodic features alone enabled to obtain more than 70% accuracy. In the combined model, audio features have, in average 2.3 more weight than textual ones.

The IM condition classification task exhibited mixed results. Only lexical features allowed to improve the results from the benchmark by more than 3 points.

In line with possible expectations, models fitted on lexical features yielded improvements against the benchmark to predict the CBCA Score, lowering the MAE by more than one-third of its dummy value. Interestingly prosodic features also produced compelling results, despite the fact that CBCA is a rigid score almost solely based on the responses' actual lexical content.

Finally, independently of the ML model or set of predictors, the Interview performance turned out to be very poorly encompassed, and none of them achieved large improvements compared to the dummy benchmark.

VI. Analysis and discussion

As described above, one surprising result is that even though the CBCA evaluation process only considers the discourse's content, predictions based uniquely on prosodic features were quite accurate. For both prosodic and lexical features, the main criteria used by the model to assess CBCA was the length of the answer measured respectively through its time duration or its number of tokens (see subsection E, Figure 5 and Figure 4), potentially reflecting the number of details. It underlines the fact that the different categories of features might be capturing similar elements.

Several hypotheses can be considered to explain why the interview performance label results are low. On the one hand, the features might not have been significant, the models may not be sophisticated enough or more data could have been required. On the other hand, it was actually expected that the interview performance score was unrelated to the features as the interviewer was comparing the content of the interviewee's answer to a set of specific criteria to assess competencies that should are not related to such features.

We might have come to different conclusions if we treated the interview performance as a classification task instead of regression. The performance was rated on a scale of 1 to

Features	Model	Education Level (acc)	IM condition (acc)	Interview performance (MAE)	CBCA Score (MAE)
	Dummy	0.53	0.35	0.68	1.64
Lexical	Ridge	0.66	0.41	*0.68*	1.20
	AdaBoost	0.65	0.37	*0.68*	1.13
	Random forest	0.65	0.38	*0.68*	1.09
Prosodic	Ridge	0.71	0.36	0.67	1.34
	AdaBoost	0.70	0.38	0.67	1.31
	Random forest	0.67	*0.35*	*0.68*	1.29
Emotions	Ridge	0.57	*0.35*	*0.68*	1.61
	AdaBoost	0.55	0.36	*0.70*	1.60
	Random forest	0.52	0.36	*0.74*	*1.68*
Lexical + Prosodic	Ridge	0.74	**0.42**	*0.68*	1.23
	AdaBoost	0.70	0.38	0.67	1.13
	Random forest	0.72	0.40	0.67	1.11
BERT-based	Albert/Small BERT	**0.95**	0.40	**0.64**	**0.94**

Table I: Models Scores

5 (discrete values only). It might also explain why for the task of predicting CBCA score, which has 28 distinct values and was also treated as a regression task, we achieved better results using the mean absolute error as a metric.

Extracted from audio, emotions were aggregated, and therefore, a lot of information was lost. More sophisticated methods are needed in order to grasp time changing aspect of an interviewee's state and relation to the presented questions. We could then expect a better result, especially for the IM condition, which requires knowing whether an interviewee lied at some point in time during their answer.

The most important coefficients of the Ridge classification for the education level are shown in Figure 2. Prosodic features are prevalent in positive coefficients. Even though they are not all easily interpretable, it can be observed that interviewees that speak louder on average and for longer were more likely to be MBA. On the other hand, lexical features are important BCom predictors, in particular confusion and the frequent use of the first person.

BERT was found to be more accurate for all tasks besides IM condition. This implies that more explainable features could be extracted from the interviews and would lead to better results.

More results can be found in the appendix in subsection E.

On a side note, we tried the prosodic features to predict gender and got a very high accuracy (0.975) with a non-optimised ridge classifier. Our models might therefore use those characteristics as a proxy for its prediction on other labels such as the education level. Those models would reproduce those biases in a production application and would require further work to make it fairer.

As the research subject is broad and there might be many improvements by adding the following models and techniques:

- checking if the speaker talks grammatically correct
- predicting emotions from text, using BERT
- checking if the candidate answers the question asked
- incorporating video-based features like gaze, smile, and emotions

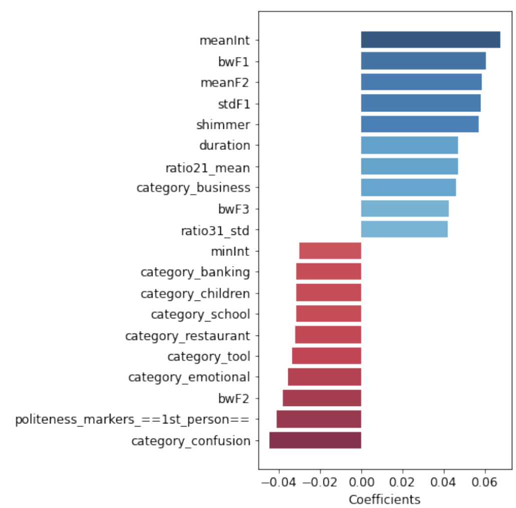


Figure 2: Lexical + Prosodic coefficients in Ridge classification for education level condition (negative coefficients associated with BCom, positive ones with MBA).

VII. Summary

It was found that using lexical and prosodic features, we were able to predict CBCA scores and the education level more accurately than our benchmark (dummy models). However, we did not reach any conclusive results for the interview performance due to the expected independence between the assessment process and the extracted features and for the IM condition, which would require to grasp state changes during the answer duration. The emotion extraction process could be extended to tackle this latest issue. BERT was also found to be more accurate for all tasks besides IM condition.

References

[1] HireVue, "Hirevue records surge in anytime, anywhere candidate engagement: Hirevue," Oct 2020. [Online]. Available: https://www.hirevue.com/press-release/hirevue-records-surge-in-anytime-anywhere-candidate-engagement

[2] N. Roulin and D. M. Powell, "Identifying applicant faking in job interviews," *Journal of Personnel Psychology*, vol. 17, no. 3, pp. 143–154, Jul. 2018. [Online]. Available: https://doi.org/10.1027/1866-5888/a000207

[3] A. Vrij and S. Mann, "Criteria-based content analysis: An empirical test of its underlying processes," *Psychology, Crime & Law*, vol. 12, no. 4, pp. 337–349, 2006.

[4] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018. [Online]. Available: http://arxiv.org/abs/1810.04805

[5] I. Naim, M. I. Tanveer, D. Gildea, and M. E. Hoque, "Automated prediction and analysis of job interview performance: The role of what you say and how you say it," in *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, vol. 1, 2015, pp. 1–6.

[6] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: Liwc and computerized text analysis methods," *Journal of Language and Social Psychology*, vol. 29, no. 1, pp. 24–54, 2010. [Online]. Available: https://doi.org/10.1177/0261927X09351676

[7] E. Fast, B. Chen, and M. S. Bernstein, "Empath: Understanding topic signals in large-scale text," *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2016.

[8] I. Montani, M. Honnibal, M. Honnibal, S. Van Landeghem, A. Boyd, H. Peters, P. O. McCann, M. Samsonov, J. Geovedi, J. O'Regan, G. Orosz, D. Altinok, S. L. Kristiansen, , Roman, Explosion Bot, L. Fiedler, G. Howard, Wannaphong Phatthiyaphaibun, Y. Tamura, S. Bozek, , Murat, M. Amery, B. Böing, P. K. Tippa, L. U. Vogelsang, B. Vanroy, R. Balakrishnan, V. Mazaev, and GregDubbin, "explosion/spacy: v3.2.0: Registered scoring functions, doc input, floret vectors and more," 2021. [Online]. Available: https://zenodo.org/record/1212303

[9] J. P. Chang, C. Chiam, L. Fu, A. Z. Wang, J. Zhang, and C. Danescu-Niculescu-Mizil, "Convokit: A toolkit for the analysis of conversations," 2020.

[10] C. Danescu-Niculescu-Mizil, M. Sudhof, D. Jurafsky, J. Leskovec, and C. Potts, "A computational approach to politeness with application to social factors," in *Proceedings of ACL*, 2013.

[11] R. Řehůřek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, May 2010, pp. 45–50.

[12] M. Röder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures," in *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*. ACM, Feb. 2015. [Online]. Available: https://doi.org/10.1145/2684822.2685324

[13] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0377042787901257

[14] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [Computer program]," Version 6.1.38, retrieved 2 January 2021 http://www.praat.org/, 2021.

[15] Y. Jadoul, B. Thompson, and B. de Boer, "Introducing Parselmouth: A Python interface to Praat," *Journal of Phonetics*, vol. 71, pp. 1–15, 2018.

[16] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PLOS ONE*, vol. 13, no. 5, pp. 1–35, 05 2018. [Online]. Available: https://doi.org/10.1371/journal.pone.0196391

[17] M. K. Pichora-Fuller and K. Dupuis, "Toronto emotional speech set (TESS)," 2020. [Online]. Available: https://doi.org/10.5683/SP2/E8H2MF

[18] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[19] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org. [Online]. Available: https://www.tensorflow.org/

[20] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 42, no. 1, pp. 80–86, 2000. [Online]. Available: http://www.jstor.org/stable/1271436

[21] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. [Online]. Available: https://doi.org/10.1023/a:1010933404324

[22] Y. Freund, R. E. Schapire *et al.*, "Experiments with a new boosting algorithm," in *icml*, vol. 96. Citeseer, 1996, pp. 148–156.

APPENDIX

A. Prosodic features

Prosodic Feature	Description
Energy	Mean spectral energy.
F0 MEAN	Mean F0 frequency.
F0 MIN	Minimum F0 frequency.
F0 MAX	Maximum F0 frequency.
F0 Range	Difference between F0 MAX and F0 MIN.
F0 SD	Standard deviation of F0.
Intensity MEAN	Mean vocal intensity.
Intensity MIN	Minimum vocal intensity .
Intensity MAX	Maximum vocal intensity .
Intensity Range	Difference between max and min intensity.
Intensity SD	Standard deviation.
F1, F2, F3 MEAN	Mean frequencies of the first 3 formants: F1, F2, and F3.
F1, F2, F3 SD	Standard deviation of F1, F2, F3.
F1, F2, F3 BW	Average bandwidth of F1, F2, F3.
F2/F1 MEAN	Mean ratio of F2 and F1.
F3/F1 MEAN	Mean ratio of F3 and F1.
F2/F1 SD	Standard deviation of F2/F1.
F3/F1 SD	Standard deviation of F3/F1.
Jitter	Irregularities in F0 frequency.
Shimmer	Irregularities in intensity.
Duration	Total interview duration.
% Unvoiced	Percentage of unvoiced region.
% Breaks	Average percentage of breaks.
maxDurPause	Duration of the longest pause.
avgDurPause	Average pause duration.

Table II: Extracted prosodic features (source: [5])

B. Additional datasets to train the emotion predictor

- RAVDESS [16]: The dataset contains 1440 audio recordings that exhibit eight emotional expressions: neutral, calm, happy, sad, angry, fearful, surprise, and disgust. Short sentences are read by 12 male and 12 female actors.
- TESS [17]: The dataset contains seven emotions. It does not include calm; therefore, we discarded it, and it contains a pleasant surprise, which will be treated the same as a surprise. Different sentences than in the RAVDESS dataset were read by two actresses in total, making 2800 recordings.

C. Audio Based emotions predictions

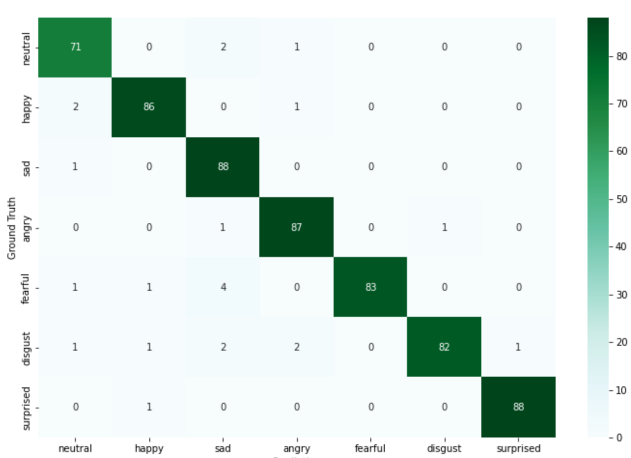


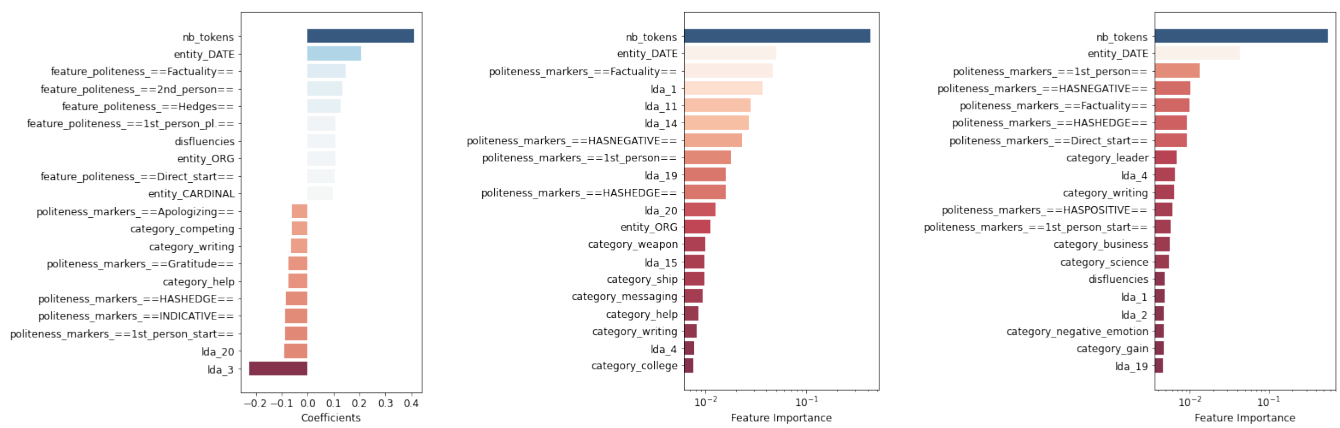
Figure 3: Confusion matrix of emotions predictions.

Features	Model	Education Level (acc)	IM condition (acc)	Interview performance (MAE)	CBCA Score (MAE)
	Dummy	0.53	0.35	0.68	1.64
Lexical full	Ridge	0.66	0.41	*0.68*	1.20
	AdaBoost	0.65	0.37	*0.68*	1.13
	Random forest	0.65	0.38	*0.68*	**1.09**
Lexical clustered	Ridge	0.62	**0.42**	**0.66**	1.17
	AdaBoost	0.60	0.40	0.67	1.17
	Random forest	0.59	0.39	*0.68*	1.17
Lexical PCA	Ridge	0.66	0.41	*0.68*	1.20
	AdaBoost	0.64	0.38	*0.70*	1.38
	Random forest	0.61	0.36	*0.70*	1.37
Prosodic	Ridge	0.71	0.36	0.67	1.34
	AdaBoost	0.70	0.38	0.67	1.31
	Random forest	0.67	*0.35*	*0.68*	1.29
Emotions	Ridge	0.57	*0.35*	*0.68*	1.61
	AdaBoost	0.55	0.36	*0.70*	1.60
	Random forest	0.52	0.36	*0.74*	*1.68*
Lexical + Prosodic	Ridge	**0.74**	0.42	*0.68*	1.23
	AdaBoost	0.70	0.38	0.67	1.13
	Random forest	0.72	0.40	0.67	1.11

Table III: Models Scores

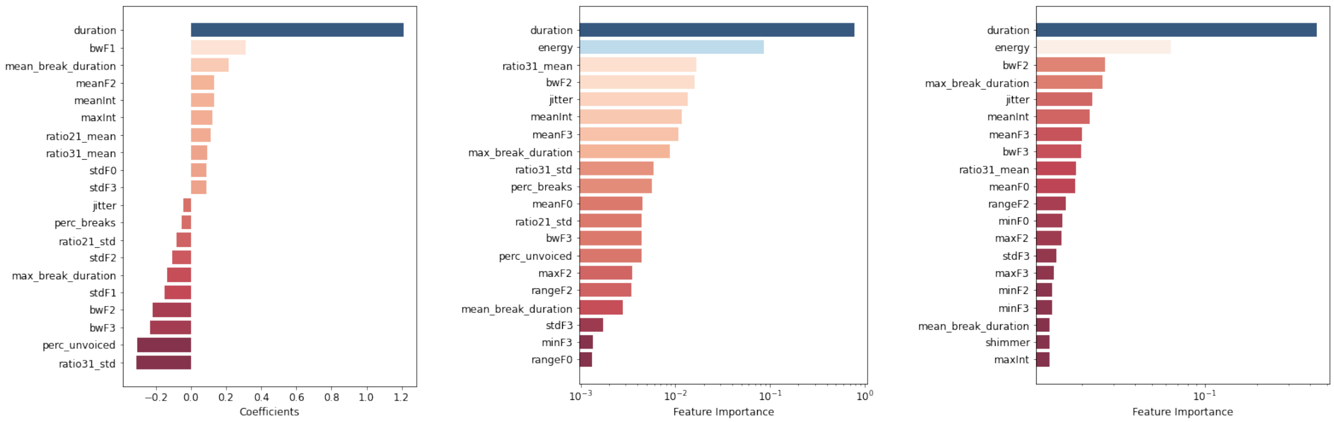
D. *Models scores*

E. *Coefficients' importance visualisations*



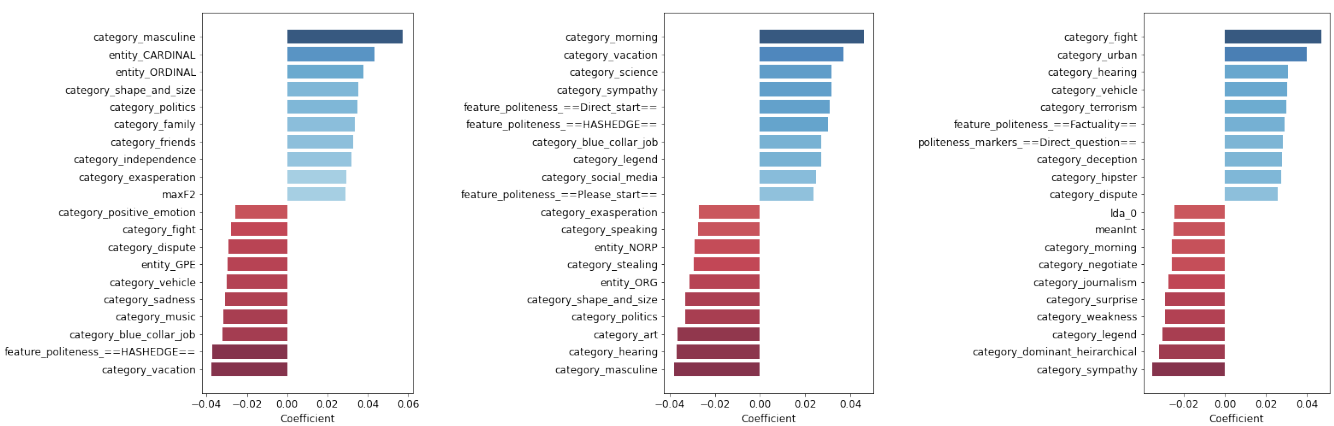
(a) Ridge. (b) ADABoost. (c) Random forest.

Figure 4: Lexical coefficients' importance for CBCA regression.



(a) Ridge. (b) ADABoost. (c) Random forest.

Figure 5: Prosodic coefficients' importance for CBCA regression.



(a) Class 1 (honest). (b) Class 2 (slight deception). (c) Class 3 (extensive deception).

Figure 6: Coefficients of the fitted one-vs-all ridge classification of the IM condition on Lexical + Prosodic features.