Using a self-attention architecture to automate valence categorization of French teenagers' free descriptions of their family relationships: a proof of concept

Mohammed Sedki[1], Nathan Vidal[2], Paul Roux[2,3]^, Caroline Barry[1]^, Mario Speranza[2,3]^, Bruno Falissard[1]^, Eric Brunet-Gouet[2,3]^

[1]CESP, Bâtiment 15-16 Inserm 1018, Hôpital Paul Brousse, Villejuif Cedex, France; [2]Université Paris-Saclay, Université Versailles Saint-Quentin-En-Yvelines, DisAP-DevPsy-CESP, INSERM UMR 1018, Le Chesnay, France; [3]Centre Hospitalier de Versailles, Service Hospitalo-Universitaire de Psychiatrie d'Adultes et d'Addictologie, Le Chesnay, France

Contributions: (I) Conception and design: E Brunet-Gouet, M Sedki; (II) Administrative support: M Speranza, B Falissard; (III) Provision of study materials or patients: C Barry, B Falissard; (IV) Collection and assembly of data: C Barry, B Falissard; (V) Data analysis and interpretation: E Brunet-Gouet, M Sedki; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Correspondence to: Mohammed Sedki, PhD. CESP, Bâtiment 15-16 Inserm 1018, Hôpital Paul Brousse, 16 avenue Paul Vaillant Couturier, 94807 Villejuif Cedex, France. Email: mohammed.sedki@universite-paris-saclay.fr.

Background: When caring for adolescents with mental health problems, family relationships need to be taken into account, whether they are considered supportive or unfavorable by the patients themselves. Recent developments in automated natural language processing (NLP) appear to offer solutions to the challenge of designing tools to evaluate the spontaneous discourse of adolescents on the subject of these relationships. This paper proposes a proof of concept of using NLP to categorize valence of family relationships described in free texts written by French teenagers. The proposed study traces the evolution of techniques for word embedding from classical categorization methods to self-attentional architectures.

Methods: After decomposing different texts in our possession into short texts composed of sentences and manual labeling, we tested different word embedding scenarios to train a multi-label classification model where a text can take several labels: labels describing the family link between the teenager and the person mentioned in the text and labels describing the teenager's relationship with them (positive/negative/neutral valence). The natural baseline for word vector representation of our texts is to build a Term Frequency-Inverse-Document-Frequency (TF-IDF) and train classical classifiers (Elasticnet logistic regression, gradient boosting, random forest, support vector classifier) after selecting a model by cross validation in each class of machine learning models. We then studied the strengths of word-vectors embeddings by an advanced language representation technique via the Bidirectional Encoder Representations from Transformers (BERT) CamemBERT transformer model, and, again, used them with classical classifiers to compare their respective performances. The last scenario consisted in augmenting the CamemBERT with output dense layers (perceptron) representing a classifier adapted to the multi-label classification and fine-tuning the CamemBERT original layers.

Results: The optimal fine-tuning depth that achieved a bias-variance trade-off was obtained by a cross-validation procedure. The results of the comparison of the three scenarios on a test dataset showed a clear improvement of the classification performances of the scenario with fine-tuning beyond the baseline and of a simple vectorization using CamemBERT without fine-tuning.

Conclusions: Despite the moderate size of the dataset and the input texts, fine-tuning to an optimal depth remains the best solution to build a classifier.

Keywords: Family relationships; adolescents; classification; word vector embedding; natural language processing (NLP)

^ ORCID: Paul Roux, 0000-0003-0321-4189; Caroline Barry, 0000-0001-6668-524X; Mario Speranza, 0000-0002-1785-5764; Bruno Falissard, 0000-0002-2418-4954; Eric Brunet-Gouet, 0000-0002-3784-7817.

Received: 20 January 2023; Accepted: 04 September 2023; Published online: 19 September 2023.
doi: 10.21037/jmai-23-8
View this article at: https://dx.doi.org/10.21037/jmai-23-8

Introduction

Research on family interactions and their perception by adolescents is a major issue for the management of patients in child psychiatry. Complex epistemological and methodological questions are raised by numerous works in the concerned disciplines, for which it seems interesting to bring new tools from artificial intelligence (AI) technologies, in particular from automated language processing. Indeed, the considerable increase in the interaction of teenagers with digital tools can be investigated with methods of analysis of verbal or textual data. In the following we confront family research on teenager populations, with the state-of-the-art natural language processing (NLP) and "sentiment analysis" (i.e., determining the emotional and subjective valence from a text) and try to produce new tools with the aim of fostering larger scale protocols. Let's note that research based on narratives is scarce and represents only 2% of the available literature in NLP applications to research on mental health conditions detection, the vast majority being based on social media database (1).

Despite the wealth of literature in family research, no consensus has been established on the variables and constructs to describe quantitatively family relationships (2). The most relevant theoretical frameworks focus on family histories (family development theory), systemic relationships between family members, family relationships with the environment, attachment relationships, social learning by children, etc. In order to study all those theoretical frameworks, Falissard and colleagues have developed a common tool between sociologists, psychoanalysts and adolescent psychiatrists to be applied to the free discourse that adolescents may hold about their own family relationships (2). A total of 194 French adolescents [age: mean ± standard deviation (SD) 14.7±2 years, 51% girls] were recruited to produce a corpus of descriptions of their family relationships (text length: 232±129 words). The instructions were: "*In the next half hour, would you please write as freely as you wish about your relationships in your family, explaining how things are. All that you write is anonymous and no parent or person from your school will read it*". These short texts were analyzed and rated by blind raters across 18 dimensions (affective environment, conflict, injustice, support, positive/negative relations, etc.) as decided by an expert consensus involving sociologists, psychoanalysts and child and adolescents psychiatrists. After a careful metrological investigation on the items, an exploratory factorial analysis was conducted and resulted in a unifactorial solution accounting for more than half the variance. This solution emphasized the positive/negative valence of relationships with other family members. Thus, it appears that relational valence constitutes a key element of the family descriptions produced by adolescents and of their mental representations of them. Falissard *et al.* also argue in favor of using this dimension as a primary endpoint in future interventional research (2). We add to their conclusion, that if the valence of relationships between individuals is a key aspect of family background, it advocates for sentiment analysis studies. To this end, NLP appears as a convenient tool to automatically analyze the analysis of adolescent free speech or writings (3).

Highlight box

Key findings
- Fine-tuned transformer models are relevant to categorize family relationship valence as well as person identities to achieve advanced semantic analysis of clinically important psychological constructs.

What is known and what is new?
- It is acknowledged that transformer models are widely used and accurate to perform sentiment analysis.
- We find that these models provide output vectors that convey both semantic information corresponding to a refined psychological construct (family relationship valence) and the identity of the persons it concerns. We also report that the best labelling performances are provided by fine-tuned BERT models ahead of classical classification methods based on TF-IDF or raw BERT output.

What is the implication, and what should change now?
- Larger-scale corpus labelling should be conducted by experts in psychology or psychiatry to provide training dataset for pretrained models like BERT to reach high prediction performances and to assess patients records or web-based datasets.

Supervised learning to analyze free texts' contents

Rating the valence of a text with supervised learning raises the challenge to find a convenient way to represent text contents so that prediction algorithms could process them. Indeed, extracting features from the text that will feed the classifiers is the starting point for using a supervised learning model. A first solution is called Term Frequency (TF), corresponding to the number of the occurrences of each word within the text, and its variant TF-Inverse-Document-Frequency (TF-IDF), which corresponds to the number of occurrences divided by the frequency of the word in the whole corpus (4,5). TF and TF-IDF have been popularized by the work of (6) in unsupervised document classification. Both can be also be used as simple text embedding methods to feed classical supervised learning models. However, TF and TF-IDF do not consider of ordered word sequences in a text and invariant to permutation of words. Taking into account word order in a text is a real challenge for improving the predictive performance of supervised learning models.

Deep-learning approaches brought a new efficient way to achieve supervised learning while accounting for word order. Relying on a general back-propagation of error mechanisms, and benefiting from large training datasets, Deep-learning approaches are now commonly used in emotion labeling tasks (7). Deep-learning methods, such as Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM) and transformer architectures, ensure a better handling of word-order, of long-term dependency, and of gradient vanishing problems. As noted by (8), deep learning methods were only recently introduced in NLP literature applied to mental health. For example, Gutiérrez *et al.* used NLP to classify texts from first-episode patients with schizophrenia compared with healthy controls based on metaphoricity assessment and sentiment analysis (emotional valence of texts) to train a Recursive Neural Network classifier (9). Some authors even advocated for the use of fine coherence and syntactic NLP processing to classify diagnosis such as psychosis (10).

Recently a method showing a qualitative leap in terms of performance for embedding textual data was proposed by Google AI via the implementation of a sophisticated neural architecture called Bidirectional Encoder Representations from Transformers (BERT) standing for BERT (11). This model implements an attentional mechanism that consists in weighting the input vectors (representing words or vector embeddings) in order to form a context vector used to process each word through successive multiple layers of nodes (i.e., a feedforward network): such models are often referred to as self-attentional architectures.

A work based on self-attentional architectures demonstrated the great performance of BERT to classify 209,000 texts from the GermEval 2020 task (12). In their study, pieces of free texts describing a person's situation, feelings and actions from simple drawings of the Operant Motive Test (13) were classified by trained psychologists into five possible motives and rated into six possible levels. Several architectures were compared such as supervised autoencoders, fully connected neural networks, and transformers [BERT, cross-lingual language models (XLMs), DistilBERT], with respect to a baseline consisting of a support vector classifier of TF-IDF text representation. Interestingly, the best classifier performance, i.e., F1-scores of 0.69, was found with a simple BERT model. Another study showed that BERT model could be successfully used to classify social media sentences into five basic emotions, with a high macro F1-score of 0.83 (14). A recent survey discussing the use of transformer-based models in mental health is proposed by (15) while (16) explores potential linguistic markers, detected by NLP methods, as a means of objectively measuring the severity of psychotic symptoms of schizophrenia in an acute clinical setting. Dai *et al.* tackled the problem of predicting 5 diagnostic classes in psychiatry (major/minor depression, bipolar disorder, schizophrenia, and dementia) from a text corpus of 500 medical records (17). Various architectural scenarios and training strategies were tested. The study's conclusions focused mainly on feasibility aspects. However, most studies trained the BERT model with larger datasets than ours (the texts of 194 adolescents). Knowing that BERT could handle small datasets, we thought it would be interesting to test its performances when trained on a small dataset, as it can be the case when studying specific populations for which data is scarce.

One question raised by the use of self-attentional models concerns the lexical, syntactic and semantic features processed by the different processing layers (namely, the BERT model is composed of 11 layers of identical structure, themselves composed of several sublayers). Since the encoder has the duty to transform N word-vectors of 768 dimensions into a single output vector of the same size, we hypothesize that each successive layer progressively reduces dimensionality while increasing abstraction. Understanding deep-learning models is generally complicated and is a research question in itself, far beyond the scope of this work. As discussed by Jain *et al.*, one would think that attentional weights are directly related to the importance given to inputs, which would help the interpretability of these models, but experimentation shows that this is not the case (18). Similarly, a close examination of BERT's attentional weights in the aforementioned GermEval classification task shows that the transformer pays more attention to form features (i.e., use of personal pronouns, stop words, negation, punctuation marks, unknown words, and some conjugation styles) than to content words (12). While interpretability of attentional weights proves difficult, other authors have conducted a layer-by-layer examination of the structure of transformers by probing the corresponding hidden outputs. In question answering tasks, it was shown that successive layers support processing allowing for named entity extraction, coreference resolution, relation classification, and supporting fact extraction (19). A layer-by-layer examination would inform if and how deep fine-tuning of pretrained models should be applied to a BERT model to achieve tasks akin to sentiment analysis.

Objectives

We will challenge as a proof-of-concept the use of NLP learning techniques to interpret the corpus of adolescent texts described in Falissard *et al.* (2). We will label text fragments based on the valence of the relationships and the person involved (e.g., mother, father, etc.). We will transform the text segments from the corpus with a BERT-derived model (11), pre-trained on a French database (20), into vectors on which classification techniques can be applied. More precisely, we will be interested in classifying the valence, like in sentiment analysis, but also the categories of the people described in the text segments (i.e., mother, father, sister, the respondent him/herself, etc.). We will compare the classification performances of the classical algorithms: elastic net logistic regression, gradient boosting classifier, random forest, support vector classifier. In addition, we will test the added value of transformers' embedding with a text vectorization method based on TF-IDF that does not take into account word order information. To go further in our understanding of machine learning usability in our field, we will test the interest of fine-tuning the upper layers of the transformer. We tackle the question of the categories of semantic information (i.e., person categories and/or relational valence) the transformer actually encodes in order to determine whether this information is represented in the output of the transformer, and usable for prediction. In addition, we raise the question of the level of fine-tuning that could improve classification performance.

Methods

In this section, we describe different steps of data preparation, classification and finally comparison of the methods. The labeled textual datasets will be common to all evaluations. We will then transform the texts into vectors either by the TF-IDF method or by applying an attentional model. Then, several families of classical classification models are used, each one having been adjusted for its main hyperparameters. We also use the possibility to extend the attentional model by a perceptron in order to obtain a prediction. The attentional model itself will be compared using different depths of fine-tuning. *Figures 1,2* illustrate the different computations used in the experiments.

Data preparation and labelling

Texts from 194 teenagers from Falissard *et al.*'s study (2) were used to generate a set of 1,648 text segments (8.4 segments in average per text; max segment length: 345 chars. This study reports an additional analysis of data collected which has obtained ethical agreement from the CCTIRS (Comité consultatif sur le traitement de l'information en matière de recherche), CNIL (Commission Nationale Informatique et Liberté) with number MG/CP 10962, conforming to the provisions of the Declaration of Helsinki (as revised in 2013). Informed consent was obtained from the participants' parents.

Note that we will use the term "segment" in the following to refer to these pieces of text). Each one was composed of one or several sentences depending on the presence of referential pronouns that have to be taken into account as a context to understand an assertion. For instance, "I have good relationships with my mother and my father with whom I live." and "I live with my mother and father. I have good relationships with them." are both taken as a single text segment to be labelled and to be processed. Thus, each segment could be unambiguously interpreted either by a human or an automated semantic analysis.

Once the segmentation was carried out, segments were labelled with 11 binary tags by one of the authors (Brunet-Gouet E) according to simple criteria on the valence (*Valence*) and type of information given and concerning the people involved in the relationship (*Subject*) described (see an example in *Table 1*). Whenever the segment contained information on relational valence, it was rated as positive (+), negative (–) or neutral (0). Positive relationships refer to a good understanding, an expression of positive affect,

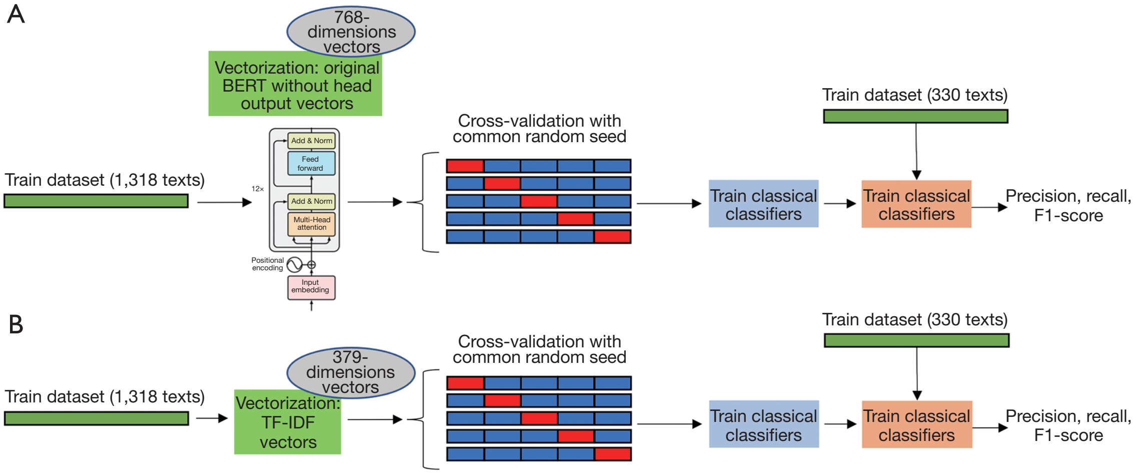
A
768-dimensions vectors
Vectorization: original BERT without head output vectors
Train dataset (1,318 texts)
Add & Norm
Feed forward
Add & Norm
Multi-Head attention
12×
Positional encoding
Input embedding
Cross-validation with common random seed
Train classical classifiers
Train dataset (330 texts)
Train classical classifiers
Precision, recall, F1-score
B
Train dataset (1,318 texts)
379-dimensions vectors
Vectorization: TF-IDF vectors
Cross-validation with common random seed
Train classical classifiers
Train dataset (330 texts)
Train classical classifiers
Precision, recall, F1-score

Figure 1 Design of the experiments. (A) Classical prediction algorithms (Elasticnet logistic regression, gradient boosting, random forest, support vector classifier) were evaluated using a 5-fold cross-validation procedure on the training dataset, with a common random seed to ensure that training and testing dataset are the same from an experiment to another. These classifiers are fed with CamemBERT's 768-dimensions vectors. (B) TF-IDF vectors are used to evaluate the classical prediction algorithms with the same cross-validation procedure based on the same random seed. BERT, Bidirectional Encoder Representations from Transformers; TF-IDF, Term Frequency-Inverse-Document-Frequency.

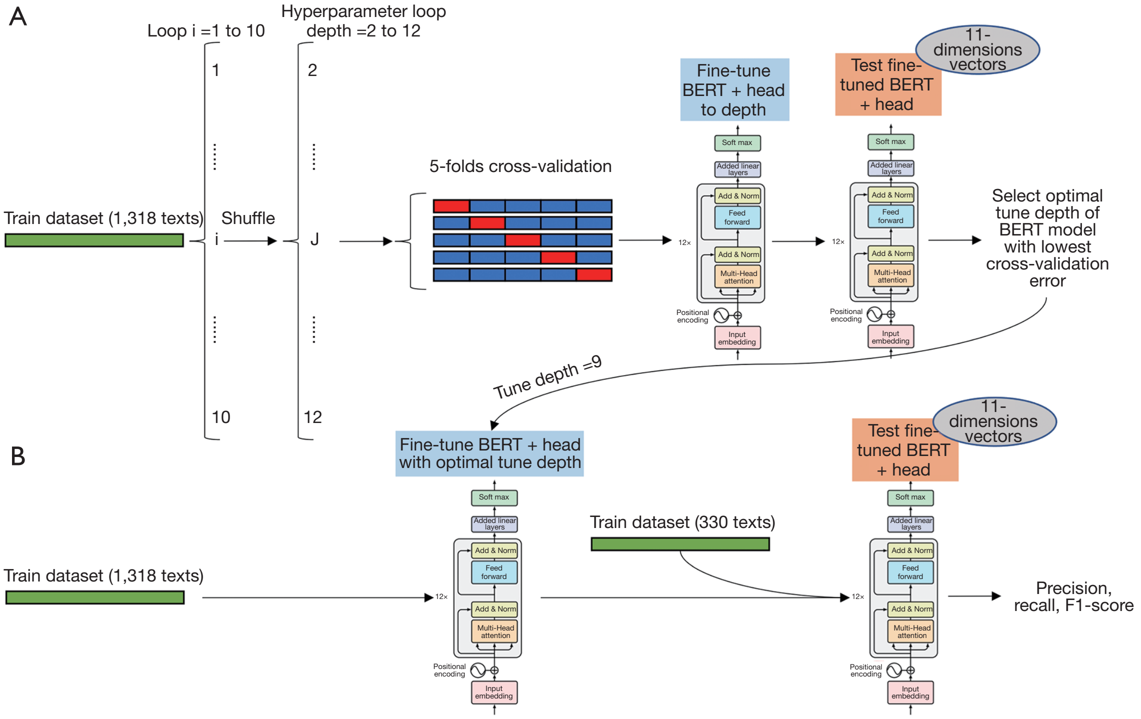
A
Loop i =1 to 10
Hyperparameter loop depth =2 to 12
1
2
Train dataset (1,318 texts)
Shuffle
i
J
10
12
5-folds cross-validation
Fine-tune BERT + head to depth
Test fine-tuned BERT + head
11-dimensions vectors
Soft max
Added linear layers
Add & Norm
Feed forward
Add & Norm
Multi-Head attention
12×
Positional encoding
Input embedding
Select optimal tune depth of BERT model with lowest cross-validation error
Tune depth =9
B
Fine-tune BERT + head with optimal tune depth
Train dataset (1,318 texts)
Train dataset (330 texts)
Test fine-tuned BERT + head
11-dimensions vectors
Precision, recall, F1-score

Figure 2 Design of the experiments. (A) To determine the optimal tune depth to fine-tune BERT and the perceptron on-top, a 5-fold cross-validation procedure, repeated ten times, was used to train the model and then measure prediction error. In subsequent cross-validation computations, this hyper-parameter was used to train CamemBERT. (B) Assessment of fine-tuned BERT was achieved on the training dataset with previously fixed learning hyper-parameters. Precision, recall and F1-sores are obtained. BERT, Bidirectional Encoder Representations from Transformers.

Table 1 Example of teenagers' description of his/her family (randomly extracted from the dataset)

Raw text	Labels										
	+	−	Neutral	Informative	Me	Brother	Sister	Father	Mother	Family	Others
Ma famille contient 4 membres : mon père, ma mère, une sœur de 10 ans et moi. (*My family consists of 4 people: my father, my mother, a 10-year-old sister and me*)				x	x		x	x	x		
Mes parents ne sont pas divorcés. (*My parents are not divorced*)				x				x	x		
Il existe des affinités plus marquées. Ma sœur ressemble beaucoup à mon père et donc s'entend mieux avec lui. Il ont le même caractère blageur et pas sérieux. (*There are more marked affinities. My sister looks a lot like my father and therefore gets on better with him. They have the same cheerful, easy-going personality*)	x						x	x			
Contrairement à moi et ma mère qui sommes beaucoup plus calme. (*Unlike me and my mother, who are much calmer*)				x	x				x		
Je ne parle que très peu à mon père, il tourne toujours tout en dérision et ne sait pas vraiment écouter. Par contre ma sœur semble lui parler avec plus de facilité (*I don't talk to my father much, he always makes fun of everything and doesn't really know how to listen. My sister, on the other hand, seems to speak to him more easily*)			x		x		x	x			
Je semble donc m'entendre bien mieux avec ma mère. (*So I seem to have a much better relationship with my mother*)	x				x				x		
Mes relations dans la famille ont donc l'air de fonctionner par pair, même si bien entendu mes parents n'ont pas de préférences marqués. Il y a rarement de réels discussions, qu'elles que soit la personne ayant des problèmes il est extrêmement rare d'en parler. (*So my family relationships seem to work in pairs, although of course my parents don't have any marked preferences. There are rarely any real discussions, and no matter who has problems, it's extremely rare to talk about them*)			x		x			x	x	x	

cooperation between the teenager and the subject (i.e., "I get along very well with my mother" or "My sister is like a close friend"). Negative relationships correspond to conflicts, disagreement, absence of a normal relationship, etc. ("My father is aggressive with me" or "My sister doesn't talk to me, she's a stranger to me"). Finally, the neutrality (*0*) of the statement is identified when the text implies an emotional relationship between the subject and another one ("I live with my mother and I see my father all the time") and/or contains both positive and negative elements (ambivalent or ambiguous feelings) and does not allow for a clear valence to be inferred ("My father is nice to me but most of the time I can't stand him"). In the absence of valence information, the text was considered as informative (*Info*) about the habits or living conditions of the persons when they did clearly imply a form of relationship (for instance, "My parents eat together in the evening with the children" do not imply that the respondent is involved and describe more a way of life than a relational involvement of the persons). The subjects described in a segment have been labeled as follows: the respondent (*Me*), "Mother", "Father", "Sister", "Brother", "Family member", "Others". As this labeling method was intended to be simple and coarse, this procedure did not require any linguistic expertise other than being proficient in the corresponding native language.

Finally, the dataset consisted of 1,648 items with their 11 labels and was randomly split into a training dataset and a test dataset of sizes 1,318 and 330, respectively. The

Table 2 Different families of classical classification methods that are compared with the corresponding list of the hyperparameters

Classical model families
Elasticnet logistic regression
Gradient boosting classifier
Random forest
Support vector classifier
Hyper-parameters
C and l_1 ratio
Learning rate and number of estimators
Maximum of features and bootstrap
C, gamma and kernel

The parameters were optimized by cross-validation for each model family. The grids of values that were tested are available in the following notebooks: https://github.com/masedki/ados_familles.

same random generator seed was used to have the same test dataset for all the compared model families.

Text vectorization

Two text vectorization methods were applied, TF-IDF and Transformers, in order to feed the classical classifiers.

Term-frequency-inverse-document

For any word *w* in a text segment *t*, TF-IDF(w, t) is the product of tf(w, t) the number of occurrences of *w* in *t*, and a weighting term $idf(w) = 1 + \log\left(\frac{1+n}{1+df(w)}\right)$ where n is the number of text samples in the whole dataset, and df(w) is the number of samples in the dataset that contains *w*. This weighting procedure dampens the impact of words that occur very frequently in a corpus which may be considered as less informative than those that occur in a small fraction of the corpus. In this work, TF-IDF transforms produced 379-dimensions vectors that were used to train and test classifiers as shown in *Figure 1B*.

Transformer models

Each segment of the corpus was first transformed into sequences of word vectors (tokenization) and then to a 768-dimensions vector, using the attentional model CamemBERT, derived from RoBERTa (21), which was trained on a corpus of French texts (20). CamemBERT was used without fine-tuning. The teenagers' lexicon was not modified beyond some typos corrections, in order to stay as close as possible to free text writing without complex preprocessing, and to evaluate attentional models robustness when taking into account this population's specific style [this approach was also used in (12)]. It is interesting to note that transformers' architectures per se do not take into account word orders. But this crucial information is taken into account by combining each word-embeddings with a corresponding positional-embedding. The output of such model is also a 768-dimensions numeric vector that may be fed to any classical classification model or into a perceptron (see *Figure 1A*).

Statistical analysis

Prediction of labels were based on two different methods: the use of classical families of classifiers fed either by TF-IDF or by BERT's output vectors as described in section "Text vectorization".

The evaluation procedure was the same for each prediction strategy including the use of cross-validation with the same random seed to ensure training and testing dataset comparability, and the use the same performance metrics. For each label, the performance of each prediction strategy was measured by three classical performance metrics in classification. These metrics are respectively: $Precision\left(Positive\ predictive\ value\right) = \frac{TP}{TP+FP}$, $Recall\left(sensitivity\right) = \frac{TP}{TP+FP}$ and $F1\text{-}score = 2 \cdot \frac{Precision.Recall}{Precision+Recall}$, where *TP* are the true positives, *FP* the false positives count. The metric Precision tells us what proportion of the positive predictions are actually positive. The metric Recall tells us what proportion of real positives is correctly classified while the metric F1-score corresponds to a measure of balance between the two previous metrics. If either the metrics Precision or Recall are low, the F1-score is low.

Classical models

Several families of classification models were compared, as listed in *Table 2*. Knowing that each family of model is based on their own sets of hyper-parameters, the best model from each one was selected thanks to a 5-folds cross validation procedure on the training dataset with features obtained either by TF-IDF (section "Data preparation and labelling") or by attention models (section "Text vectorization"). The different hyper-parameters of each family of models are listed in *Table 2*. Each pair of models optimized by the cross-

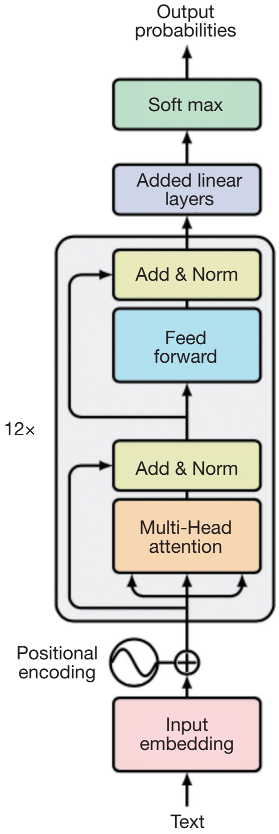

Figure 3 This network summarizes from bottom to top the transformer based neural architecture involved in the section "Transformer model fine-tuning". The block corresponding to CamemBERT is repeated 12 times while the classification network has been added by us in order to carry out the multi-label classification task. The set of weights of the added layers as well as the weights of the last three layers of multi-headed attention at the top of the CamemBERT block are learned on the training set. Indeed, tune depth involving three transformer layer blocks was determined by a repeated cross-validation procedure. BERT, Bidirectional Encoder Representations from Transformers.

validation procedure for each family was finally evaluated on the test dataset. The hyper-parameters optimization and the final model fit on the training dataset was done using *scikit-learn* library (22) excepting gradient boosting classifier which uses *lightgbm* library. *Figure 1A,1B* illustrate these computations.

Transformer model fine-tuning

The comparison of the set of classifiers listed in section "Classical models" follows two steps. The first step consists in transforming a text segment into a feature vector using the procedures described in sections "Term-frequency-inverse-document" and "Transformer models" and a step of choosing and training the model to predict the labels vector from the feature vector obtained in the previous step. In this section, we focus on a model for predicting labels from text in a single step. This was possible by extending CamemBERT with a classification perceptron placed at its' output.

On the top of CamemBERT model, a 3-layers perceptron was added in order to predict the 11 labels. An encoding layer actually encompasses several neural layers including self-attention and *feed-forward* networks. The last layer of the transformer corresponding to the CLS (*Classification*) token was composed of 768 units, that were progressively reduced to 200, 110 and 11 units, using three perceptron layers placed at the *head* of the transformer. Nonlinear *hyperbolic tangent* activation function was used for the first two layers (i.e., 768 units to 200, and 200 units to 110), and, for multiple labeling, the last layer (i.e., 110 units to 11). *Figure 3* schematizes this architecture.

A binary cross-entropy loss function (*torch.nn.BCEWithLogitsLoss()*) was used for training in this multi-label-multi-output situation where more than one labels may be found in a single text segment. The weights and the biases of 3-layer perceptron were always back-propagated. The number of transformer's encoding layers that were fine-tuned corresponds to tune depth hyper-parameter which was selected using a 5-fold cross-validation procedure which was globally repeated ten times in order to reduce possible variability in the results. The remaining embedding layers below the tune depth layer were frozen during the training procedure. The procedure allowing to determine the optimal tune depth is schematized in *Figure 2A*, and testing of this model with metrics similar to the one used to assess classical classifiers is illustrated in *Figure 2B*.

Results

Fine tuning

Figure 4 represents curve of 5-fold cross validation error repeated ten times as a function of the hyper-parameter tune depth. The repetition of the cross-validation procedure was used to remove random effects that would occur during the 5-fold cross validations. Other hyper-parameters were also fixed as described in the notebooks https://github.com/masedki/ados_familles. The main result from this procedure was that best performances were found when the

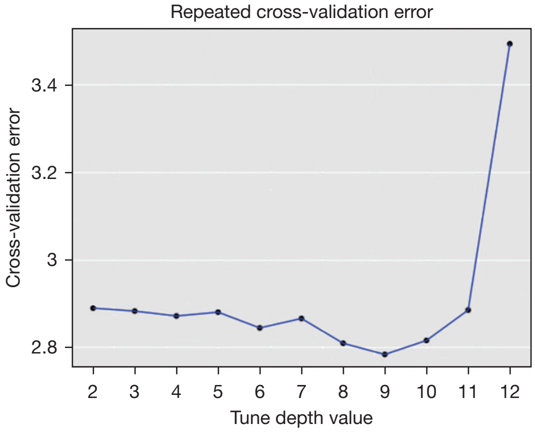

Figure 4 Cross-validation error rate (binary cross-entropy error) as a function of the hyper-parameter tune depth during fine tuning. Horizontal axis corresponds to the lower layer to which layer parameters are back-propagated from the output. Twelve corresponds to fine tuning the perceptron only while BERT's parameters are frozen. BERT, Bidirectional Encoder Representations from Transformers.

transformer's parameters were updated from the ninth layer to the perceptron's output. It is also worth noting that only training the perceptron on top of the transformer penalizes significantly the error rate. Please note that optimization of tune depth required 1 week of processing with a NVIDIA A100 Tensor Core 40 GB GPU.

Classification performances

Table 3 summarizes all the results of the numerical computations presented in section "Statistical analysis". Overall, the strategy consisting in fine-tuning CamemBERT provides the best, yet far from perfect, performances in the majority of the labels. Indeed, the highest F1-scores are systematically found with the Fine-tuning procedure. Concerning valence labels, the best classification performances of fine-tuned transformers are found for positive (+) labels (0.74) in comparison with classical classifiers trained either on BERT output vectors or TF-IDF vectors (F1-scores respectively from 0.51 to 0.62, and 0.53 to 0.57). It appears that the positive label is associated with higher recall (0.82) than precision (0.67), while the inverse pattern is found for the negative label (–). Neutrality labels (0) were very poorly identified, demonstrating the ambiguity inherent in their definition. The Informative label (*Info*) was associated with high precision (0.78), recall (0.81) and F1-scores (0.79). Obviously, labels that convey people identity are correctly classified with metrics most of the time superior to 0.9, a fact that demonstrates the relative simplicity of recognizing these entities. The label indicating a reference to the respondent himself (*Me*) is generally well identified, whatever the classifier used. However, the results show that some classical classifiers provide only poor predictive capability for certain labels such as *Brother*, or *Others*.

Discussion

In the present work, we aimed at testing the feasibility of using automatic language processing methods based on transformers to categorize the writings of French adolescents on their family relationships. Based on previous research, we considered that relational valence could constitute a relevant element as a psychological outcome. We used transformers because these models are pre-trained on large corpora allowing us to benefit from the "general" linguistic and semantic knowledge encoded inside and to fine-tune them on smaller datasets of labeled sentences. We wanted to see if a model recently made available in French would have sufficient semantic representation capacity to determine valence, as in sentiment analysis, as well as to identify the people described in the texts. To begin with the technical aspects associated with learning the 11 labels, the hyper-parameters selected are consistent with published studies employing BERT or its variants.

In this study, we find that models based on fine-tuning the inner layers of the transformer outperform those based on classifying the output vectors of the transformer head. Other works have also reported that intervening in the internal structure of BERT could have a benefit, although this strategy is debatable for that it separates the new finetuned model from the original one. The question was whether there is a level of depth to which the backpropagation of the error must access to maximize performance. A Study on BioBERT model to classify multiple clinical concepts has shown that freezing up to six bottom layers of the encoder during training maintained good performances (23). In the present work, best error rate over validation set could be found around the ninth layer. It can be suggested that the learning depth corresponding to better prediction performances informs us about the type of information that are processed along the different layers of the transformer. Van Aken *et al.* raise the hypothesis that different types of processing and representation exist

Table 3 Comparison of different types of error on the test dataset

	+	−	Neutral	Informative	Me	Brother	Sister	Father	Mother	Family	Others
Support	72	63	47	146	263	36	57	124	123	72	23
Precision											
Fine-tuning	0.67	0.85	0.40	0.78	0.86	0.91	0.97	0.97	0.94	0.91	0.78
Elasticnet lr	0.65 (0.69)	0.71 (0.75)	0.28 (0.33)	0.74 (0.61)	0.85 (0.81)	0.74 (0.79)	0.89 (0.98)	0.90 (0.93)	0.83 (0.87)	0.77 (0.81)	0.67 (0.75)
Gradient boosting	0.73 (0.65)	0.88 (0.44)	0.25 (0.45)	0.75 (0.64)	0.84 (0.80)	0.75 (0.78)	0.88 (0.92)	0.84 (0.93)	0.79 (0.85)	0.83 (0.73)	0 (0.83)
Random forest	0.82 (0.70)	0.71 (0.73)	0 (0.33)	0.71 (0.64)	0.82 (0.80)	0 (0.72)	0.89 (0.92)	0.86 (0.93)	0.76 (0.89)	0.73 (0.77)	0 (0.67)
SVC	0.68 (0.70)	0.79 (1)	0 (0)	0.76 (0.66)	0.85 (0.81)	0.68 (0.79)	0.91 (0.94)	0.90 (0.92)	0.83 (0.89)	0.82 (0.79)	0.50 (0.79)
Recall											
Fine-tuning	0.82	0.37	0.38	0.81	0,97	0.89	0.98	0.93	0.93	0.86	0.61
Elasticnet lr	0.56 (0.49)	0.40 (0.10)	0.11 (0.02)	0.73 (0.75)	0.94 (1)	0.39 (0.64)	0.68 (0.81)	0.76 (0.85)	0.82 (0.85)	0.67 (0.64)	0.26 (0.52)
Gradient boosting	0.49 (0.49)	0.22 (0.11)	0.02 (0.11)	0.75 (0.66)	0.97 (0.95)	0.17 (0.81)	0.49 (0.86)	0.70 (0.85)	0.69 (0.90)	0.47 (0.64)	0.00 (0.43)
Random forest	0.38 (0.43)	0.08 (0.17)	0 (0.04)	0.75 (0.59)	0.96 (0.92)	0.00 (0.86)	0.30 (0.95)	0.52 (0.92)	0.60 (0.93)	0.33 (0.71)	0.00 (0.52)
SVC	0.57 (0.46)	0.30 (0.05)	0 (0)	0.77 (0.73)	0.94 (1.00)	0.42 (0.64)	0.72 (0.86)	0.77 (0.78)	0.81 (0.89)	0.64 (0.58)	0.09 (0.48)
F1-score											
Fine-tuning	0.74	0.51	0.39	0.79	0.91	0.90	0.97	0.95	0.93	0.89	0.68
Elasticnet lr	0.60 (0.57)	0.51 (0.17)	0.15 (0.04)	0.74 (0.67)	0.89 (0.89)	0.51 (0.71)	0.77 (0.88)	0.82 (0.89)	0.83 (0.86)	0.72 (0.71)	0.38 (0.62)
Gradient boosting	0.58 (0.56)	0.35 (0.18)	0.04 (0.17)	0.75 (0.65)	0.90 (0.87)	0.27 (0.79)	0.63 (0.89)	0.77 (0.89)	0.74 (0.88)	0.60 (0.68)	0 (0.57)
Random forest	0.51 (0.53)	0.14 (0.28)	0 (0.08)	0.73 (0.61)	0.89 (0.86)	0.00 (0.78)	0.45 (0.93)	0.65 (0.92)	0.67 (0.91)	0.46 (0.74)	0 (0.59)
SVC	0.62 (0.55)	0.44 (0.09)	0 (0)	0.77 (0.69)	0.90 (0.89)	0.52 (0.71)	0.80 (0.90)	0.83 (0.84)	0.82 (0.89)	0.72 (0.67)	0.15 (0.59)

The performance of the one-step learning technique described in *Figure 2B* corresponds to the row labeled fine-tuning. The lines entitled Elasticnet lr, gradient boosting, random forest and SVC correspond to the learning methods that we have labeled as classical and described in section "Classical models". The metrics given in brackets correspond to the TF-IDF text vectorization scenario described in *Figure 1B* while the metrics given outside the brackets correspond to the text vectoring scenario by attention model described in *Figure 1A*. BERT embeddings' results are written in plain text, tests of classification performances using the same algorithms on TF-IDF sentence vectorization are reported between parentheses. lr, logistic regression; SVC, support vector classifier; TF-IDF, Term Frequency-Inverse-Document-Frequency; BERT, Bidirectional Encoder Representations from Transformers.

within transformer networks and conclude that "it could be beneficial to fit parts of the network to specific tasks in pre-training, instead of using an end-to-end language model task" (19). Although our approach (selecting the depth of fine-tuning) differs technically from theirs (selecting the output of an internal layer of the transformer), we concur in the idea that these pre-trained models have an internal architecture whose knowledge would help optimizing new tasks.

The predictive performance of the fine-tuned model concerning either the relational semantics of sentences or the identity of persons was compared with that of classical classification algorithms. The first result is that the last layer of BERT conveys information about the identity of individuals at higher performances (i.e., larger F1-scores) by the support vector classifier compared to Elasticnet, gradient boosting and random forest. Moreover, BERT fine-tuning, as described above, brings a substantial gain on precision and recall compared to classical algorithms. Fine-tuned BERT is also found better when compared with classical classifiers fed with TF-IDF vectors instead of word-vectors embeddings. This suggests that transformers have a sufficient power of representation of subjects, in terms of categories of people, to deal with relational situations. As such, this result did not solve a real methodological issue and does not impose this technology as efficient to detect these entities because simple parsing algorithms can easily do so. But it shows that transformers have the capability of making it possible to embed the information of the agents being represented in the texts in combination with other semantic information for any purpose and learning.

Now regarding semantic labels representing valence, we obtain contrasted prediction performances. Compared with classical algorithms, the best F1-scores were obtained using BERT fine-tuning for the +, –, 0 and "info" labels (i.e., positive, negative, neutral and information labels). However, only + (positive) and "info" labels were reached 0.7 F1-scores, with a better Recall than Precision measures. As it stands, the proposed finetuning method, on a small sample of texts, presents a better recall capacity for positive and informative texts. This approach may be used for research in large corpora of texts and aiming at extracting the maximum number of texts dealing with positive relations or even distinguishing them from informative texts. However, it is interesting to note that the – (negative) label is associated with a better precision. This can also have the advantage, in large text corpora, of targeting selectively negative texts with a reduced number of false positives. In any case, if the precision/recall profiles of the labels turn out to be distinct, it seems necessary to conduct additional investigations to see if their use should be thought in distinct scenarios of use.

Limitations of the study

In this study we used a small learning base and tried to use it to train sophisticated NLP models. These models being pre-trained on very large textual databases were to benefit from their ability to represent relevant semantic information. The unfavorable results concerning the recall of the negative valence labels are possibly related to the small size of the training dataset. We conclude that a substantial effort to build larger databases of labels realized by human operators could help progressing. Concerning the labeling of complex psychopathological criteria by experts, this may lead to quite expensive work. If such databases are created, the possibilities of translating them automatically from one language to another or of using multilingual models will have to be evaluated.

Conclusions

In this paper, we propose the use of recent NLP methods in child and adolescent psychiatry. To our knowledge, no study has previously investigated the possibility of labeling texts related to family relationships using these methods. Our work brings contrasted preliminary results notably by showing that labels concerning positive relational valence are predicted with better precision and recall. On the other hand, negative valences are more complicated to label and our results do not provide an immediate solution to detect difficult family situations. Further work should improve the model in order to meet the requirements of clinical use. Nevertheless, the use of these methods, despite their limited predictive power, should be considered for large-scale investigations of internet and social media databases to characterize the evolution of young people's views of their family relationships. Finally, we concur with Abbe and colleagues, on the idea that new text mining techniques might discover new variables from the clinical experiences reported directly by the patients (24). Beyond "classical applications" such as diagnosis or suicide prediction, one could propose automatizing of validated clinical measures or psychological constructs. However, to achieve these goals, important efforts to constitute properly labelled text corpus would be necessary with the perspective of training large language models.

Acknowledgments

We wish to warmly thank Christine Hassler (INSERM U669, Université Paris-Sud and Université Paris-Descartes, Paris, France), Muriel Letrait (CERLIS, Université Paris-Descartes, Paris, France), Guillaume Macher (CERLIS), François Marty (EA 4046, Université Paris-Descartes, Paris, France), Elsa Ramos (CERLIS), Anne Revah-Lévy (INSERM U669), Philippe Robert (EA 4046), and François de Singly (CERLIS), for conceiving the multidisciplinary protocol to gather the adolescents' texts dataset used in this study and published in reference (2).
Funding: None.

Footnote

Data Sharing Statement: Available at https://jmai.amegroups.

com/article/view/10.21037/jmai-23-8/dss

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://jmai.amegroups.com/article/view/10.21037/jmai-23-8/coif). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. This study reports an additional analysis of data collected and published in Reference 2 which has obtained ethical agreement from the CCTIRS (Comité consultatif sur le traitement de l'information en matière de recherche), CNIL (Commission Nationale Informatique et Liberté) with number MG/CP 10962, conforming to the provisions of the Declaration of Helsinki (as revised in 2013). Informed consent was obtained from the participants' parents.

References

1. Zhang T, Schoene AM, Ji S, et al. Natural language processing applied to mental illness detection: a narrative review. NPJ Digit Med 2022;5:46.
2. Falissard B, Barry C, Hassler C, et al. When Assessing Intra-Familial Relationships, Are Sociologists, Psychoanalysts and Psychiatrists Really Considering Different Constructs? An Empirical Study. PLoS One 2015;10:e0132153.
3. Liu B. Sentiment Analysis: Mining Opinions, Sentiments, and Emotions. London: Cambridge University Press; 2015.
4. Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. Inf Process Manag 1988;24:513-23.
5. Salton G, McGill MJ. Introduction to Modern Information Retrieval. New York, NY: McGraw-Hill, Inc.; 1986.
6. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. Journal of Machine Learning Research 2003;3:993-1022.
7. Alswaidan N, Menai MEB. A survey of state-of-the-art approaches for emotion recognition in text. Knowledge and Information Systems 2020;62:2937-87.
8. Le Glaz A, Haralambous Y, Kim-Dufor DH, et al. Machine Learning and Natural Language Processing in Mental Health: Systematic Review. J Med Internet Res 2021;23:e15708.
9. Gutiérrez ED, Cecchi G, Corcoran C, et al. Using Automated Metaphor Identification to Aid in Detection and Prediction of First-Episode Schizophrenia. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark: Association for Computational Linguistics; 2017:2923-30.
10. Corcoran CM, Mittal VA, Bearden CE, et al. Language as a biomarker for psychosis: A natural language processing approach. Schizophr Res 2020;226:158-66.
11. Devlin J, Chang MW, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics; 2019:4171-86.
12. Villatoro-Tello E, Parida S, Kumar S, et al. Applying Attention-Based Models for Detecting Cognitive Processes and Mental Health Conditions. Cognit Comput 2021;13:1154-71.
13. Kuhl J. Auswertungsmanual für den Operanten Multi-Motiv-Test OMT (IMPART Test-Manuale), 1999. Available online: https://impart.de/produkt/auswertungsmanual
14. Nijhawan T, Attigeri G, Ananthakrishna T. Stress detection using natural language processing and machine learning over social interactions. Journal of Big Data 2022;9:33.
15. Greco CM, Simeri A, Tagarelli A, et al. Transformerbased language models for mental health issues: A survey. Pattern Recognition Letters 2023;167:204-11.
16. Jeong L, Lee M, Eyre B, et al. Exploring the Use of Natural Language Processing for Objective Assessment of Disorganized Speech in Schizophrenia. Psychiatr Res Clin Pract 2023. doi: 10.1176/appi.prcp.20230003.
17. Dai HJ, Su CH, Lee YQ, et al. Deep Learning-Based Natural Language Processing for Screening Psychiatric Patients. Front Psychiatry 2020;11:533949.
18. Jain S, Wallace BC. Attention is not Explanation. In:

Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics; 2019:3543-56.

19. van Aken B, Winter B, Löser A, et al. How Does BERT Answer Questions?: A Layer-Wise Analysis of Transformer Representations. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management. Beijing, China: Association for Computing Machinery; 2019:1823-32.
20. Martin L, Muller B, Ortiz Suárez PJ, et al. CamemBERT: a Tasty French Language Model. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics; 2020:7203-19.
21. Liu Y, Ott M, Goyal N, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. 2019. doi: 10.48550/arXiv.1907.11692.
22. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 2011;12:2825-30.
23. Kalyan KS, Sangeetha S. BertMCN: Mapping colloquial phrases to standard medical concepts using BERT and highway network. Artif Intell Med 2021;112:102008.
24. Abbe A, Grouin C, Zweigenbaum P, et al. Text mining applications in psychiatry: a systematic literature review. Int J Methods Psychiatr Res 2016;25:86-100.

doi: 10.21037/jmai-23-8

Cite this article as: Sedki M, Vidal N, Roux P, Barry C, Speranza M, Falissard B, Brunet-Gouet E. Using a self-attention architecture to automate valence categorization of French teenagers' free descriptions of their family relationships: a proof of concept. J Med Artif Intell 2023;6:15.