



**HAL**  
open science

# **Interdisciplinary approach to identify language markers for post-traumatic stress disorder using machine learning and deep learning**

Robin Quillivic, Frédérique Gayraud, Yann Auxéméry, Laurent Vanni, Denis Peschanski, Francis Eustache, Jacques Dayan, Salma Mesmoudi

## **► To cite this version:**

Robin Quillivic, Frédérique Gayraud, Yann Auxéméry, Laurent Vanni, Denis Peschanski, et al.. Interdisciplinary approach to identify language markers for post-traumatic stress disorder using machine learning and deep learning. *Scientific Reports*, 2024, 14 (1), pp.12468. <10.1038/s41598-024-61557-7>. <hal-04595510>

**HAL Id: hal-04595510**

**<https://hal.science/hal-04595510v1>**

Submitted on 17 Jul 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.




Distributed under a Creative Commons CC BY 4.0 - Attribution - International License



OPEN

## Interdisciplinary approach to identify language markers for post-traumatic stress disorder using machine learning and deep learning

Robin Quillivic<sup>1,2</sup>, Frédérique Gayraud<sup>3</sup>, Yann Auxéméry<sup>4,5</sup>, Laurent Vanni<sup>6</sup>, Denis Peschanski<sup>7,8</sup>, Francis Eustache<sup>1,9,10</sup>, Jacques Dayan<sup>1,9,10,11</sup> & Salma Mesmoudi<sup>1,2,7,8</sup>

Post-traumatic stress disorder (PTSD) lacks clear biomarkers in clinical practice. Language as a potential diagnostic biomarker for PTSD is investigated in this study. We analyze an original cohort of 148 individuals exposed to the November 13, 2015, terrorist attacks in Paris. The interviews, conducted 5–11 months after the event, include individuals from similar socioeconomic backgrounds exposed to the same incident, responding to identical questions and using uniform PTSD measures. Using this dataset to collect nuanced insights that might be clinically relevant, we propose a three-step interdisciplinary methodology that integrates expertise from psychiatry, linguistics, and the Natural Language Processing (NLP) community to examine the relationship between language and PTSD. The first step assesses a clinical psychiatrist's ability to diagnose PTSD using interview transcription alone. The second step uses statistical analysis and machine learning models to create language features based on psycholinguistic hypotheses and evaluate their predictive strength. The third step is the application of a hypothesis-free deep learning approach to the classification of PTSD in our cohort. Results show that the clinical psychiatrist achieved a diagnosis of PTSD with an AUC of 0.72. This is comparable to a gold standard questionnaire (Area Under Curve (AUC)  $\approx$  0.80). The machine learning model achieved a diagnostic AUC of 0.69. The deep learning approach achieved an AUC of 0.64. An examination of model error informs our discussion. Importantly, the study controls for confounding factors, establishes associations between language and DSM-5 subsymptoms, and integrates automated methods with qualitative analysis. This study provides a direct and methodologically robust description of the relationship between PTSD and language. Our work lays the groundwork for advancing early and accurate diagnosis and using linguistic markers to assess the effectiveness of pharmacological treatments and psychotherapies.

There is evidence to suggest that language may play a role in the development and maintenance of PTSD. Studies using qualitative methods have shown that individuals with PTSD often have difficulty expressing their thoughts and feelings about the traumatic event and may use language in ways that are different from people without the disorder<sup>1–5</sup>. Moreover, many of the psychotherapies recommended involve language from the defusing and debriefing stages onwards<sup>6</sup>. For example, a diagnosis scale based only on language markers (the SPLIT), has recently been proposed<sup>7</sup>. However, to the best of our knowledge, no study evaluates the diagnosis power of language on PTSD.

PTSD is a frequent endemic mental health condition that might develop after a person experiences or witnesses a traumatic event (Criterion A of DSM-5). In this work, we based our analysis on the symptoms of PTSD

<sup>1</sup>PSL-EPHE, Paris, France. <sup>2</sup>ISCFIP, Institut des Systèmes Complexes, Paris île-de-France, France. <sup>3</sup>Laboratoire dynamique du langage, UMR 5596, CNRS, université Lyon-II, Lyon, France. <sup>4</sup>Centre Hospitalier de Jury-les-Metz, centre de réhabilitation pour adultes, Metz, France. <sup>5</sup>UMR 1319 Inspiire, INSERM, Université de Lorraine, 9 avenue de la forêt de Haye, Nancy, France. <sup>6</sup>CNRS, UMR 7320 : Bases, Corpus, Langage, Nice, France. <sup>7</sup>Université PARIS 1 Panthéon-Sorbonne, Paris, France. <sup>8</sup>CNRS, CESSP, UMR 8209, Paris, France. <sup>9</sup>INSERM, NIMH U1077, Caen, France. <sup>10</sup>UNICAEN, Caen, France. <sup>11</sup>CHU de Rennes, Rennes, France. ✉email: robin.quillivic@ephe.psl.eu

as defined by the DSM-5 (American Psychiatric Association. 2013). These symptoms are: being exposed to a traumatic event (criterion A of DSM-5), reliving the traumatic event (criterion B), avoiding reminders of the event (criterion C), negative changes in cognition and mood (criterion D), changes in physical and emotional reactions (criterion E), and the creation of distress or functional impairment in a person's life (criterion G).

Establishing a diagnosis of PTSD and its symptoms has always been challenging in clinical practice owing to the clinical characteristics of the disorder itself, in particular the difficulty in confiding this experience, which is symptomatic of the cognitive and behavioral avoidance strategies for anything that might remind the patient of the trauma. On the one hand, some criteria may also fluctuate over time, creating “partial” PTSD in the nosographic sense, but just as disturbing for the patient. This is the reason why researchers have proposed and studied<sup>18,9</sup> Partial PTSD (P-PTSD): recent studies analyzing the data from the attacks in Paris on the night of the 13th November 2015 also include P-PTSD in their cohorts<sup>10,11</sup>. On the other hand, comorbidities such as anxiety, depression, addiction, and social maladjustment, are often better identified than post-traumatic symptoms, which perpetuate a poor overall prognosis. The medico-economic consequences of increased morbidity and mortality in untreated patients are major<sup>12</sup>. Although effective pharmacological and psychotherapeutic interventions exist, and considerable progress has been made over the past two decades in shedding light on the biological effects, particularly through functional brain imaging, the absence of clinically available biomarkers remains a challenge. Can language analysis improve post-traumatic disorder diagnosis in practice?

Recent research in Natural Language Processing (NLP) has demonstrated that NLP can provide indicators of psychopathology, particularly for Depression, PTSD, Suicide, and Psychosis<sup>13–15</sup>. These Machine Learning (ML) or Deep Learning (DL) models might identify risk characteristics using spoken or written communications resulting in low-cost and low-effort healthcare systems<sup>14</sup>. This trend is possibly confirmed by the fivefold increase in the number of publications on mental illness detection using machine learning or deep learning methods over the last 6 years<sup>16</sup>. In spite of this tendency, there are no unified datasets or gold-standard methods to compare publications. Authors use different diagnostic tools and data sources to build NLP models that often lack interpretability<sup>17</sup>.

Social media posts from Twitter (X) and Reddit associated with self-declared diagnosis are often used<sup>18–20</sup>, which enables the use of large datasets with a huge number of individuals but lacks homogeneity. Moreover, studies often compare exposed cohorts with PTSD to non-exposed cohorts without PTSD. Therefore, it is difficult to know if the linguistic markers are related to exposure to psychic trauma or to the diagnosis of chronic PTSD itself. Clinical notes<sup>21</sup> and non-structural interviews<sup>22–24</sup> are also used and often associated with a more precise diagnosis using self-questionnaire PCL-5 based on the DSM-5 criterion<sup>25</sup> or semi-structured interview SCID (American Psychiatric Association 2013). To build NLP models, many kinds of linguistic features are extracted: statistical (number of words, number of words per sentence), morpho-syntactic (proportion of first-person pronoun, verb tense), topic modeling (LDA, LSA); word vector representation (Word2Vec, Doc2Vec, Glove, Fasttext), contextual embeddings vectors (BERT, Roberta), graph-based features<sup>26</sup>, coherence<sup>27</sup> and readability features<sup>28</sup>, external resources such as LIWC<sup>29</sup>, sentiment analysis scores like LabMT<sup>30</sup>, TexBlob (Loria, 2018) or FLAIR<sup>31</sup> and transfer learning methods like DLATK<sup>32</sup> that used pre-trained models on social media data. The models used for the classification task, which consists of separating in people with and without PTSD, are mainly Random Forest (RF)<sup>33</sup>, Logistic Regression (LR), CNN, LSTM, and transformers<sup>15,16</sup>.

The NLP and qualitative analysis seem to converge on a few language markers that characterize PTSD: (1) an overuse of first-person singular pronouns (I, “je”), (2) an underuse of third-person generic pronouns (it, “on”), (3) a greater number of words related to depression, anxiety, and death, (4) an overuse of spatial and temporal cues, (5) more negative emotions<sup>19,34</sup>.

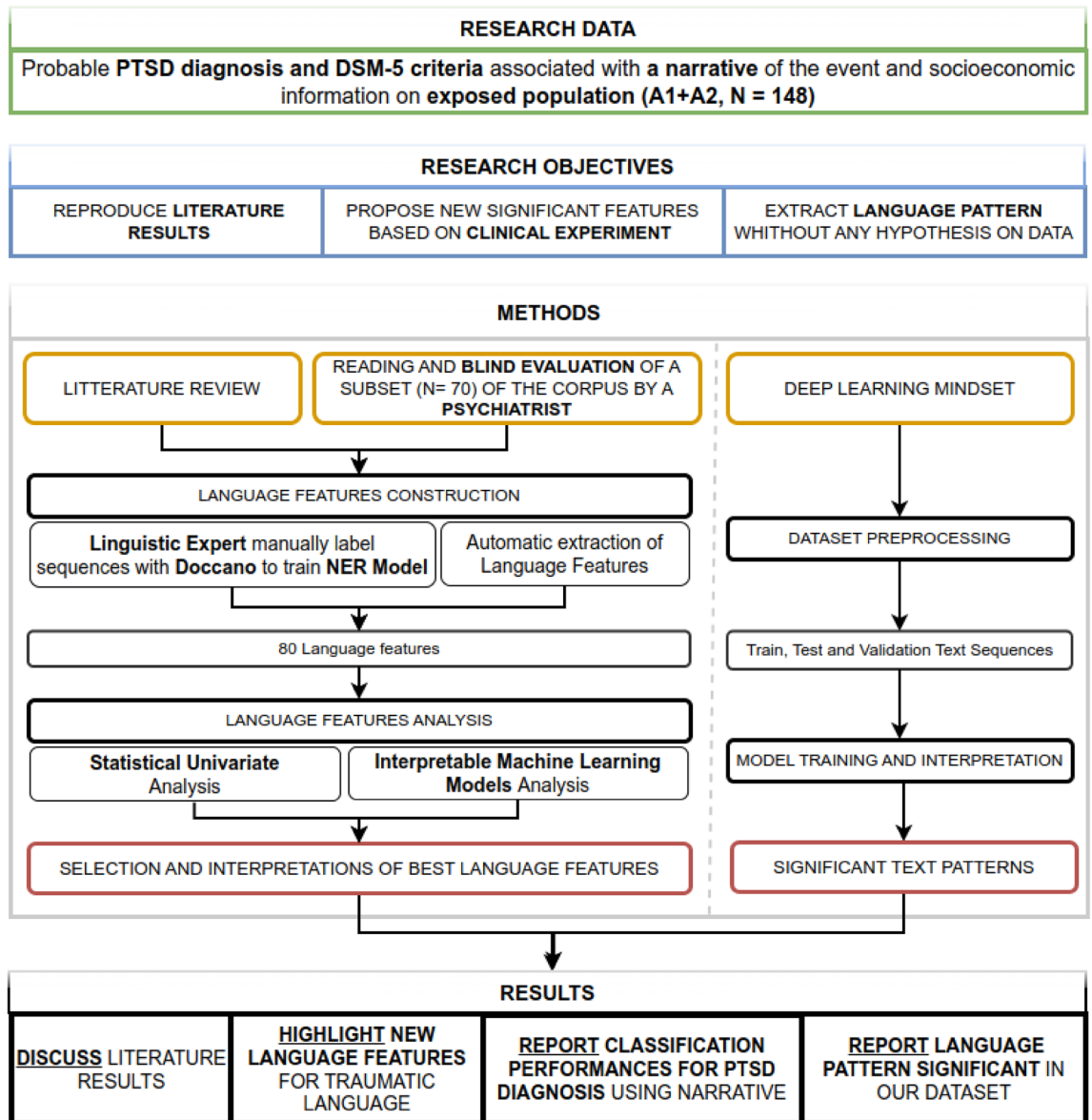
However, although the results seem compelling, these techniques have not been translated into clinical decision support systems<sup>35</sup> because of their lack of interpretability, transparency, and generalizability<sup>36</sup> (the degree to which the results can be applied to a broader context). This study is a first step towards tackling these challenges by presenting an explainable and transparent NLP-based PTSD assessment algorithm from incorporating language features. Our contributions are the following:

1. We report the first data on how well a psychiatrist can infer PTSD and symptoms solely from reading transcripts.
2. We propose an interdisciplinary workflow (see Figs. 1 and 2) using different modeling mindsets: Statistical Frequentist Approach to describe associations, Explainable Machine Learning to decrypt the interaction among the features and measure the inference capabilities, and Interpretable Deep Learning to explore new language patterns.
3. We describe the link between language and PTSD using an original dataset that enables us to describe all the PTSD's DSM-5 criteria in a homogenous population.
4. We have designed new methods and adapted existing ones for language feature extraction for French corpora and implemented them in Python. The implementation is accessible and re-usable on another dataset here (see Supplementary Table S0 in Supplementary Material).

## Methods

### Data

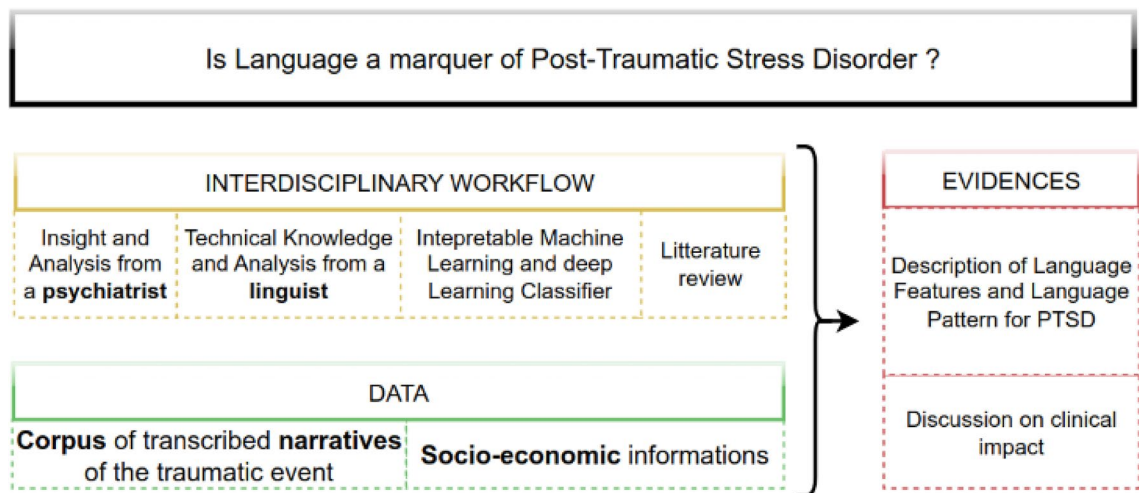
On the 13th of November 2015, three coordinated terrorist attacks occurred in Paris and its suburbs. The targets were the Bataclan concert hall, nearby trendy cafes in the 10th and 11th districts of Paris, and the *Stade de France* in a close suburb of Paris. It caused the death of 130 people, among which 90 were in the Bataclan, and



**Figure 1.** Graphical abstract that summarizes our methodological approach and contributions (A1: Direct Exposure, A2: Witnessing Trauma).

it physically wounded 413 individuals. The data used in this study were collected by the “Etude 1000” protocol which includes collecting the narratives of 934 volunteers who were at varying distances from the attacks (from survivors to ordinary citizens of a provincial town). The particular and innovative methodology is based on three questionnaires submitted to the volunteers (cf Supplementary Fig. S1 in SM), most of which were videotaped. Moreover, this longitudinal study will take place four times, in 2016, 2018, 2021, and 2026. The present analysis focuses on the data collected in 2016<sup>37,38</sup>. In our research, we particularly focused on untrained exposed people, thus, we selected only the participants that verified the DSM-5 criterion A1 (directly experiencing the traumatic event) or A2 (witnessing or having been threatened at the time of the attacks). Table 1 provides a socio-demographic description of this particular sub-sample (N = 148).

Both active and passive recruitment methods were used (see “Recruitments and Training” section in SM). Participants were recruited voluntarily, through information given by newspapers, the city hall of Paris, and associations of victims. Active recruitment involved contacting potentially eligible individuals by the program team. Participants were included from April 1st, 2016 to November 10th, 2016<sup>1</sup>, 5–11 months after the traumatic event. Filmed interviews took place in the studios of either the French National Audiovisual Institute (INA) or the Defense Communication and Audiovisual Production Agency (ECPAD). The film interview, divided into two parts, collects data about emotional, physical, and social reactions. In the first part of the interview, the participants were asked to produce a narrative relating to their experience of the attack (Q1) and also produced analysis on causes and consequences (Q2). It was followed by a shorter structured interview, an emotional memory questionnaire, that included 16 closed and some open-ended questions about post-traumatic symptoms and behaviors according to DSM-5<sup>39</sup>. In this study, we manually selected the answer to the first question of the



**Figure 2.** Description of the research mindset adopted in this article. The evidence presented is the result of an interdisciplinary workflow applied to an original dataset.

Variable	A1	A2	p value	SE	Power
N	110	38	–	–	–
% Female	55.5	68.4	0.22	0.10	0.22
Age (year), mean ± std	37 ± 9	42 ± 12	0.01**	0.04	0.93
Students (%)	3.6	7.9	0.03*	0.3	0.86
Retired (%)	0.9	0			
Unemployed (%)	7.3	2.6			
Employee (%)	8.2	2.6			
Intermediate Profession (%)	20.9	15.8			
Executive and higher profession (%)	48.2	52.6			
Worker (%)	0.9	0	0.42	0.14	0.28
Craftsman and merchant (%)	3.6	18.4			
Master degree or higher	59.1	57.9			
Bachelor degree or equivalent	27.3	21.1			
High School diploma or less	11.8	21.1			
Single (%)	61.8	63.2	0.96	0.005	0.05
Living alone (%)	67.3	68.4	0.93	0.003	0.05

**Table 1..** Socio-economic information of our cohort and statistical association with criterion A. In our cohort, criterion A is reduced to (A1: direct exposure, A2: witnessing the trauma)

semi-structured interview: “To begin with, could you tell me about November 13, 2015?”, we truncated the narrative of the traumatic event after the person had returned home or was taken to the hospital. We provided how informed consent was collected in the supplementary information.

PTSD diagnosis was assessed by professional psychiatrists specialized in PTSD using the answers from a semi-standardized trauma interview that enabled us to retrieve probable DSM-5 criterion, provided in SM of this study. The semi-standardized trauma interview was built by professional psychiatrists and its metrological qualities were assessed by comparing it with the results of the Structured Clinical Interview for DSM-5 (SCID) from a subsample (N = 85/934, see Supplementary Table S4). Participants were diagnosed with PTSD in its partial form (n = 42) if they had re-experiencing symptoms (criterion B), that caused significant distress and functional impairment (criterion G).

The interviews were transcribed using the Vocapia (Speech to Text API | Vocapia) software and then corrected by a human annotator. The corpus we examine is made up of 148 interviews and counts 849 725 words. On average, an interview is 5553 words long ( $\mp$  3628 words).

### Ethical statement

Informed consent was obtained from all subjects, details on the procedure is presented in SM. The “Etude 1000”, component of the Program, was agreed upon by the “Comité d’évaluation éthique Inserm-CNRS (IRB) n°16321

which approved all experimental protocols. All methods were carried out in accordance with the EU general data protection regulation (GDPR).

### Informed consent statement

Before they participated in this study, all participants were provided with a detailed explanation of the research objectives, procedures, potential risks, benefits, and their rights as participants. All participants provided their voluntary informed consent by signing a consent form indicating their understanding of the study and their willingness to participate. It was explicitly communicated that participation was entirely voluntary, and participants were assured that they could withdraw from the study at any time without facing any consequences. Their confidentiality and privacy rights were emphasized, and they were made aware of how their data would be collected, stored, and used. The informed consent form is available upon request from the corresponding author.

### Language features (LF)

From literature and expert knowledge (see Table 11), we formulated assumptions concerning the aspects of language that might characterize PTSD and associated symptoms. We then translated these hypotheses into quantitative measures that we extracted using human annotation and common NLP resources. We call these measures Language Features (LF) and regroup them in Table 10 according to the language aspect they concern. For instance, the proportion of present tense verbs in a narrative is a morphosyntactic Language Feature. A detailed and exhaustive list of all LF is available in the supplementary information, Supplementary Table S1 with their respective implementations.

- Sentiments and Emotion Features (n = 27), there is a rich literature exploring the link between emotion-related words or scores and PTSD narratives. Negative emotions are often associated with PTSD<sup>17,18,20,40</sup>. We used several external resources to compute scores associated with emotions, the French LIWC<sup>41</sup>, composed of 64 categories, from which we selected only 4 categories (liwc\_death, liwc\_body, liwc\_positive\_emotion, liwc\_negative\_emotion), the EMA database<sup>42</sup>, is composed of 10 categories, the FEEL lexicon<sup>43</sup> composed of 6 categories (fear, anger, surprise, sadness, disgust, and joy), the EMPATH lexicon<sup>44</sup> composed of 36 semantical and emotional categories and finally, the LabMT dictionary<sup>30</sup> translated into French that constructs a score of happiness. We also computed the Textblob<sup>45</sup> polarity and subjectivity scores (Tutorial: Quickstart—TextBlob 0.16.0 documentation).
- Lexical Features, (n = 4) were extracted using a fine-tuned token classifier based on the CamemBERT<sup>46</sup> architecture using the spacy-transformer template from spaCy. We constructed our dataset by manually labeling the documents using Doccano (<https://github.com/doccano/doccano>), an open-source tool for collaborative human annotation. The annotations were made by an experienced linguist. The models were trained using 200-word length sequences, 70% of the data were used for the training, 20% to test and fine-tune hyperparameters, and 10% to evaluate the performances of the model. In order to ensure correct performance evaluation, the evaluation sequences were taken from documents that were not used during the training. In Supplementary Table S3, we describe the dataset used and the performances for each task.
  - Lexical fields of death, body, physical sensations, and perceptions: using LIWC<sup>29</sup>, a previous study found that a larger number of sensory/perceptual words in trauma narratives was associated with PTSD<sup>22,47</sup>. However, as pointed out by the authors, LIWC has several limitations and might be easily wronged by idiomatic phrases. In our dataset, in the sentence: “Ils sont venus de derrière”, (“They came from behind”), the word “derrière” (behind) was associated by LIWC to the body but should have been associated with spatial words. In the same process as before we trained custom NER models for these categories: death, body, verbs of sensory perception, words of sensory perception, and physical sensation (see Supplementary Table S3).
- Morphosyntactic Features (n = 18), the proportion of pronouns and verb tenses were extracted using the stanza python library<sup>48</sup>, and custom NER models were trained using the same procedure as described in lexical features, in order to extract the values of present tense and third personal pronoun. See Supplementary Table S3, for performances and training data information.
  - The proportion of Pronouns and verb Tenses, which proposed a model trained on the French-gsd (GitHub—UniversalDependencies/UD\_French-GSD) dataset that enables us to retrieve the Part-Of-Speech and Morphological characteristics. Therefore we computed the proportion of present, past, future, and conditional tenses for each narrative, as well as the proportion of first (singular and plural), second, and third personal pronouns.
  - Values of the present tense: Based on the literature review<sup>49</sup>, several studies<sup>40,50,51</sup> observed an increased use of the present tense among the narrative of PTSD groups. The morpho-syntactic features will enable us to confirm this observation on our dataset. However, the present tense might have different meanings depending on the context. It can have a historical value, referring to the past, and it also makes the speech more alive. Another meaning is generic, to express general truths like definitions or properties. Finally, it can have an enunciation value by referring to the present moment, to describe an ongoing action. These different values of the present tense can only be differentiated by the context. This is the reason why models based on contextual embedding should be relevant to differentiate them<sup>52</sup>.
  - Values of “on” pronoun: As present tense, some pronouns can have different meanings according to their context; the generic pronoun plays an important role in trauma narratives<sup>1</sup>. In our study, we investigate the different values of the “on” pronoun. It can be used as “we”, for example: “On est entré au Bataclan à 20h45” (“We entered the Bataclan at 8:45 pm”). But it can also be used as a synonym for ‘someone’:

“On m’a marché dessus” (“Someone stepped on me”). Finally, it can be used generically: “on n’est jamais mieux servi que par que par soi même” (“you are never better served than by yourself”).

- Syntactic Features:
  - Passive Voice Features (n = 4), in the literature, the use of passive voice was associated with Mild Depression<sup>53</sup> but its association with PTSD has not been described so far. Yet, the grammatical passive voice is a marker of agency which is widely studied in the trauma literature<sup>54</sup>. No tools were available to assess the passive voice in a French text, so we developed our tool, using an English tool<sup>55</sup> as a model. Based on the linguistic literature in spoken French, we developed a rules-based algorithm and manually annotated 2530 passive voices from 25 narratives of our dataset. The performances are reported in Supplementary Table S2.
- Speech Disfluency Features (n = 9), the literature analyzing the phonetic data, shows that language disfluencies (fillers, repetition, hesitations, repairs, false starts, prolongation, etc.) are associated with psychosis<sup>56</sup> and PTSD<sup>57</sup>. In this study we cannot work on speech data; however, it was possible to retrieve some disfluencies from the transcription such as the repetition of syllables, the false starts, hesitation, and silent breaks marked by “...” after a word or between 2 words (score\_disfluencies), and the fillers by using a list of the most used in French (“euh”, ah, bah, etc.). All the features were normalized by the length of each testimony.
- Readability Features (n = 5), measures assessing the linguistic complexity or the readability of spoken or written productions have been widely applied to mental health detection<sup>58,59</sup> but rarely to PTSD characterization. In our study, we use the readability toolbox<sup>60</sup>, see Supplementary Table S1 for the details of features.
- Graph Structure Narrative Features (n = 12), were extracted following the methods presented in Mota<sup>67</sup>, which represented textual data in graphs and using graph-theoretical tools that might be able to capture specific features of the flow of thought. As presented in supplementary materials, connectivity measures, transitivity, and the number of loops in 1,2,3 nodes, etc. were computed for each narrative.

## Analysis

### *Psychiatrist’s blind analysis*

An initial assessment of the study involved a psychiatrist with no prior knowledge of the participants. From the entire dataset, we carefully handpicked 70 cases that are representative in terms of exposure and diagnosis, which accounted for approximately 50% of the global dataset. The clinician analyzed the narratives of the traumatic events, evaluating each criterion and making an independent diagnosis. In instances where no conclusive evidence was found, we opted to consider it as absent. The purpose of this section of the study is not to evaluate the psychiatrist’s effectiveness in diagnosing PTSD, but rather to examine the amount of information presented in a transcription that can be used to identify PTSD symptoms.

Statistical analysis, was conducted to describe demographic, clinical, and text features between two (with and without probable PTSD; with and without each symptom taken separately) or three groups (probable PTSD, probable partial PTSD, no PTSD). In the first case, we used a Mann–Whitney U test<sup>61</sup>, and in the second, we used a one-way analysis of variance (ANOVA). We followed the literature’s recommendations<sup>62</sup> concerning the Statistical Power (P) and Effect Size (ES). We used a chi<sup>2</sup>-contingency statistical test for categorical features (sexe, profession etc.)

Machine Learning (ML) analysis, was conducted to complete the previous statistical analysis that cannot capture complex relations between our features.

1. Model selection: To provide a comprehensive analysis, we chose three different models interpretable by design, thus enabling us to retrieve the critical features. The hyperparameters of the models were selected using grid search algorithm (more information in the experimental report presented in the SI).
  - Logistic Regression (LR), is a commonly used model for classification problems due to its simplicity and model interpretability. To avoid overfitting, we used elasticnet regularization (l1\_ratio = 0.6 and C = 0.1).
  - Random Forest (RF), which is an ensemble algorithm based on decision trees<sup>33</sup>. Using low-correlated weak models can produce ensemble predictions with high accuracy. We choose parameters to avoid overfitting on a small dataset. (n\_estimators = 40, min\_sample\_split = 0.4, min\_samples\_leaf = 15)
  - Explainable Boosting Machine (EBM<sup>63</sup>), is a glass box model, designed to have accuracy comparable to state-of-the-art machine learning methods like Random Forest and Boosted Trees while being highly intelligible and explainable. EBM is a generalized additive model (GAM) with few major improvements. First, EBM learns each feature function, using modern machine-learning techniques such as bagging and gradient boosting. The boosting procedure is carefully restricted to training on one feature at a time in a round-robin fashion using a very low learning rate so that feature order does not matter. Second, EBM can automatically detect and include pairwise interaction terms. The parameters were the following: max\_leaves = 10, min\_samples\_leaf = 15, max\_bins = 20 and early\_stopping\_rounds = 20.
2. Features selections. To avoid overfitting and training difficulty we reduced the number of features. Among the sentiment and emotional features, many were colinear. Hence, we only kept the sentiment features based on the emotional valence, which enabled us to keep only 6 sentiment features (Textblob, feel\_positive, labMT, and liwc\_emo\*). The features concerning the passive voices were redundant, we only kept the normalized count of the passive voice. With these 2 rules, we kept for ML analysis only 55 features out of the 80 built. In SI, we evaluate the impact of the removed features on classification performance (Supplementary Fig. E1).
3. Training procedure. We split our data set into training and testing datasets (train: 80%, test: 20%), using stratified sampling on the exposition (criterion A from DSM-5) and probable diagnosis. We used the training

dataset to select the best training configuration. We augmented the training size using the synthetic minority oversampling technique (SMOTE<sup>64</sup>) because we have an under-represented label (label: no PTSD). Then we computed the receiver operating characteristic (ROC) scores on the test dataset for 3 interpretable ML classifiers. The final scores were averaged over 100 separate randomized runs to demonstrate the robustness and stability of the results.

4. Interpretation. The models chosen were easily interpretable, we selected the models that achieved the best average performances over 100 random runs and we averaged the feature importances over these 100 runs.
5. Error analysis. On the same principle, we identified the 20 documents that are, on average, the most misclassified by the best models. We conducted statistical analysis on this subgroup to identify their specificity and enhance our model interpretations.

#### Deep learning analysis

To complete our analysis, we propose a sub-study based exclusively on a deep learning mindset that enables us to extract language patterns without apriori on our dataset. Indeed, the feature engineering described previously is particularly effective on small-size datasets but comes with many assumptions and hypotheses.

We employed a Convolutional Neural Network (CNN) model to solve our binary classification problem. To address the class imbalance in our dataset, we used a focal loss function<sup>65</sup> with  $\alpha = 0.7$  and  $\gamma = 2$ . The CNN model had an embedding size of 128, the vocabulary size was fixed to 2000, a hidden size of 32, and a kernel size of 9. To prevent overfitting, we employed several regularization techniques, including spatial dropout, L2 regularization, and early stopping. We trained the CNN classifier on a maximum of 50 epochs using annotated data from a training set (80%), and a validation set (20% of the training set) was used to monitor training. The performances were computed on a separate testing set (20%). The performance evaluation used the ROC-AUC scores averaged over 100 random runs.

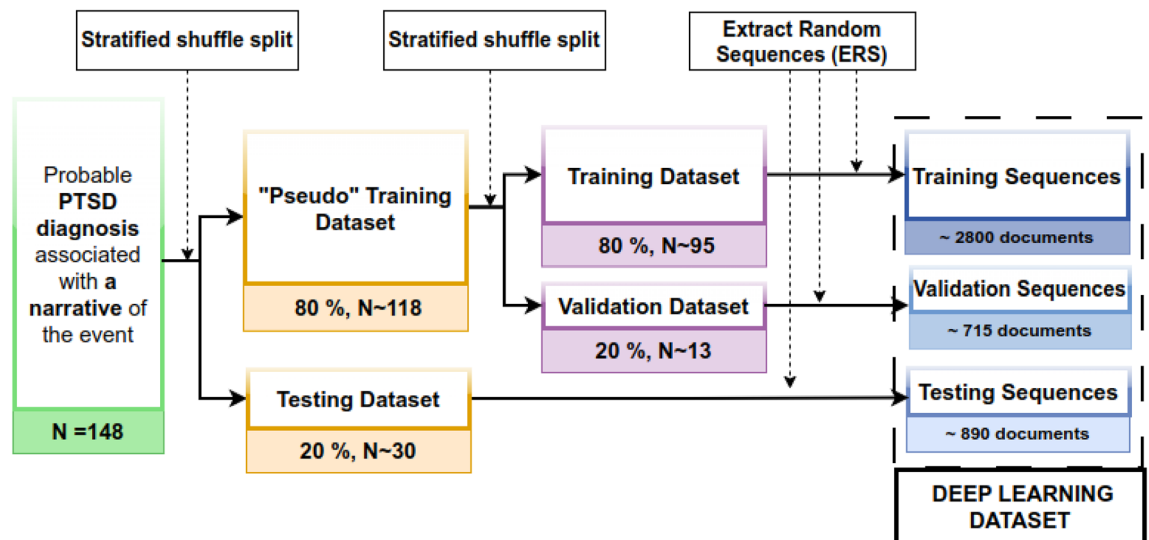
To interpret the CNN classifier's predictions, we used two techniques: Class Activation Mapping (CAM) and Deconvolution. CAM highlights the important regions of the input that contribute to the classification decision<sup>66</sup>, while Deconvolution visualizes the contribution of each input feature to the final classification score<sup>67</sup>. We followed the methodology proposed by<sup>68</sup> to apply these techniques to textual data. These techniques provide insight into how the CNN classifier makes its predictions and aids in interpreting the model's output. To emphasize the interpretability of the pattern extracted, we also computed and interpreted a multi-channel text CNN<sup>69</sup> that produces a pattern composed of a lemma, full form, and Part-Of-Speech tags.

The CNN models require the input to have the same length. We randomly extracted  $K (= 30)$  sequences of 512 tokens from each document.  $K$  was defined as the length of the 25% longest documents divided by the sequence length (512). The split into training and testing datasets was performed before the cutting of the sequence, avoiding data leaks. This process, as described in Fig. 3, is sensitive to random seeds and is slightly different over each training. The inference process described in Fig. 4 shows how we use the model to infer probable diagnosis on an unseen narrative by averaging the probability of all sequences.

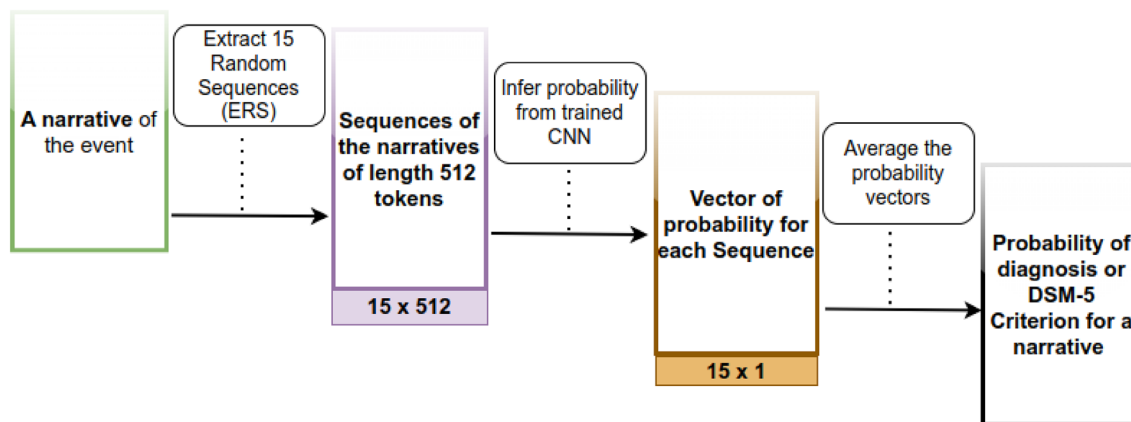
Finally, the pattern extracted using the TDS scores will be qualitatively analyzed to propose future research directions.

## Results

Our objective was to examine the potential of language and narrative structure as indicators of Post-Traumatic Stress Disorder (PTSD). We conducted three complementary studies: Firstly, we employed a blind evaluation of a subset of the corpora by a psychiatrist, who specialized in psychotraumatology, which enabled us to evaluate



**Figure 3.** Description of the process of building the Deep Learning Dataset to avoid data leakage and ensure representative distribution.



**Figure 4.** Description of the Inference process for a new document, using the CNN model.

how possible and easy it is to classify PTSD symptoms based on transcripts. Secondly, we sought to replicate the psychiatrist's evaluation through an extensive statistical analysis of text and narrative features, complemented by a highly interpretable machine learning approach, to characterize PTSD. Lastly, we introduced a hypothesis-free approach using a deep learning (text Convolutional Neural Network) classifier to identify potential language patterns that may have been missed by the previous methods. Supplementary Table S5 summarizes the results of the various approaches.

We will initially present the results of the univariate statistical analysis conducted on socio-demographic and psychopathological characteristics within our dataset. Subsequently, we will discuss the evaluation of the blind analysis made by a clinician. Moving forward, we will dive into the analysis of language features by presenting the key findings of the univariate statistical analysis (complete tables are provided as SM), followed by an examination of the performance and interpretation of the machine learning classifier. Finally, we will provide a comprehensive description of the text CNN classification method and its corresponding interpretations.

### Demographic, psychopathological, and linguistic characteristics

As demonstrated in Table 1, only age is associated with criterion A and PTSD diagnosis. Except for this feature, there is no statistical association between criterion A and socio-economic information, nor, see Table 2, with PTSD diagnosis. The length of the interview is slightly associated with the diagnostic ( $p$ -value = 0.05,  $EF$  = 0.03, power = 0.97) but strongly associated with the exposition type ( $p$ -value =  $5.10 \times 10^{-6}$ ,  $EF$  = 0.12, power = 0.99): People who were directly exposed to the traumatic event produced longer narratives (see Figure S4 and S5).

### Psychiatric blind analysis

Our first approach to evaluate the link between Language and PTSD was made by assessing the ability of a psychiatrist to blindly evaluate the DSM-5 criterion and PTSD diagnosis based only on the transcription. Table 3,

Variable	PTSD	Partial PTSD	No PTSD	p value	SE	Power
N	70	42	36	–	–	–
% Female	56.5	71.4	51.4	0.11	0.17	0.46
Age (year), mean $\pm$ std	36 $\pm$ 9	37 $\pm$ 10	42 $\pm$ 10	0.01	0.05	0.99
Students (%)	7.2	4.8	0	0.07	0.29	0.6
Retired (%)	0	0	5.4			
Unemployed (%)	10.1	2.4	0			
Employee (%)	5.8	11.9	2.7			
Intermediate Profession (%)	18.8	14.3	27.0			
Executive and higher profession (%)	44.9	52.4	54.1			
Worker (%)	0	2.4	0			
Craftsman and retailer (%)	5.8	7.1	10.8	0.58	0.13	0.17
Master degree or higher	62.3	50.0	62.2			
Bachelor degree or equivalent	21.7	33	24.3			
High School diploma or less	13.0	16.7	13.5			
Single (%)	67.7	67.7	48.6	0.14	0.17	0.4
Living alone (%)	25.6	30.9	49.6	0.04*	0.2	0.61

**Table 2.** Socio-economic information of our cohort and statistical association with PTSD.

	Partial or Full PTSD	Criterion B (intrusion sx)	Criterion C (avoidance sx)	Criterion D (negative changes in cognition and mood)	Criterion E (Hyperarousal sx)	Criterion G (Functional significance sx)
Precision	0.76	0.88	0.53	0.80	0.90	0.85
Recall	0.77	0.71	0.53	0.77	0.88	0.63
Specificity	0.65	0.72	0.59	0.65	0.71	0.78
AUC score	0.71	0.71	0.56	0.71	0.80	0.71

**Table 3.** Evaluation of the blinded rating of a human expert (clinician psychiatrist).

highlights that the global diagnosis is inferred with relatively high precision and recall, as it achieves 0.72 ROC-AUC scores. With the exception of criterion C, which is poorly identified, the other criteria are identified with high precision and recall.

### Language features statistical analysis (see Supplementary Table S1 in SM)

- Full or Partial PTSD is significantly (Table 4) associated with more death-related (DEATH, liwc\_death), body-related, and physical-sensations words (model\_BODY, model\_PHYSICAL\_SENSATIONS), less lexical diversity (token\_ratio\_score, noum\_ratio\_score), more repetitive discourse (L2, L3 degree\_average, PE, L3, average\_clustering), longer narratives (words\_number, sentence\_number), higher use of passive voice (passive\_count\_norm), more disfluencies (disfluencies\_score), and finally fewer positive emotions in narratives (labMT\_happiness\_score, gobin\_positive\_score, etc.).

Textual features	p value	SE (cohen)	Power
Words_number	1.10 <sup>-2</sup>	- 0.5	0.7
Sentence_number	2.10 <sup>-2</sup>	- 0.4	0.6
Significant emotional features			
LabMT_score	1.10 <sup>-2</sup>	0.6	0.9
Feel_positive_score	7.10 <sup>-3</sup>	0.6	0.9
Liwc_positive_emotion_score	9.10 <sup>-3</sup>	0.5	0.7
Gobin_positive_score	1.10 <sup>-2</sup>	0.5	0.8
Gobin_joy_score	1.10 <sup>-2</sup>	0.6	0.9
Significant lexical features			
Model_DEATH	2.10 <sup>-4</sup>	- 0.7	0.9
Model_BODY	2.10 <sup>-3</sup>	- 0.6	0.9
Model_PHYSICAL_SENSATIONS	9.10 <sup>-4</sup>	- 0.6	0.9
Liwc_death	1.10 <sup>-2</sup>	- 0.3	0.4
Significant morphosyntactic features			
Model_GENERIC_PRESENT	3.10 <sup>-2</sup>	0.5	0.7
Significant syntactic features			
Passive_count_norm	4.10 <sup>-2</sup>	- 0.4	0.6
Significant speech disfluencies features			
Disfluencies_score	5.10 <sup>-2</sup>	- 0.4	0.67
Significant readability features			
Token_ratio_score	1.10 <sup>-2</sup>	0.5	0.8
Noum_ratio_score	2.10 <sup>-2</sup>	0.5	0.7
Significant graph discourse features			
Graph_L3	1.10 <sup>-2</sup>	- 0.3	0.4
Graph_L2	1.10 <sup>-2</sup>	- 0.4	0.5
Graph_PE	2.10 <sup>-3</sup>	- 0.4	0.7
Graph_avarage_shortes_path	2.10 <sup>-2</sup>	0.5	0.9
Graph_average_clustering	5.10 <sup>-2</sup>	- 0.5	0.7
Graph_degree_average	1.10 <sup>-2</sup>	- 0.5	0.7

**Table 4.** The significant associations between Language Features and probable full or partial PTSD diagnosis, using Mann-Whitney U statistical test.

- Intrusions Symptoms (criterion B), are associated (see Table 5) with significantly more death, and physical sensations vocabulary and the number of disfluencies is positively associated with the reviviscence symptoms.
- Avoidance Symptoms (criterion C) are only significantly associated with more body words ( $p = 3.10^{-3}$ ,  $SE = -0.4$ ,  $power = 0.7$ ) more physical sensations words ( $p = 0.04$ ,  $SE = -0.4$ ,  $power = 0.7$ ), a higher proportion of auxiliaries in the narratives ( $p = 0.01$ ,  $SE = -0.5$ ,  $power = 0.7$ ) and higher use of past participles ( $p = 3.10^{-3}$ ,  $SE = -0.5$ ,  $power = 0.8$ ).
- Negative changes in cognition and mood (criterion D) are significantly associated (see Table 6) with more death words, both the LIWC category and our custom model, a higher proportion of “on” pronouns used with someone’s meaning (impersonally/generically), a higher proportion of pronouns in the narrative, a higher use of the first personal pronoun plural (we).
- Hypervigilance Symptoms (criterion E) are associated (see Table 7) with higher use of the passive voice, higher use of death and body related words, a reduced lexical diversity (noun\_token\_ratio, adverb\_ratio\_score), less polarity (textblob\_polarity), fewer positive emotions (labMT, gobin\_joy, feel\_positive) and fewer negative emotions (feel\_sadness, empath\_angry, empath\_fear, gobin\_sadness, feel\_anger), more repetitions (PE, L2, degree\_average, average\_shrotest\_path g0), lesser use of logical connectors (score\_generical\_connector\_matches), more future tenses (verb\_indicatif\_future) and longer narratives (words\_number, sentences\_numbers).
- Functional significance symptoms (criterion G), are associated (see Table 8) with longer narratives, higher use of the passive voice, higher use of death, physical sensations, and body-related words, less lexical diversity (token\_ratio\_score), more repetitive narratives according to graph measures (PE, L2, degree average), more first personal pronouns (I, me), and less emotional polarity (labMT score, gobin\_positive)

### Machine learning analysis

Model Performance best performances were mainly achieved using Logistic Regression with ElasticNet regularization. Performances average on 100 random runs on the test dataset for each classification task and each model is presented in Table 9. The global diagnosis ( $AUC = 0.69 \pm 0.09$ ), the criterion D ( $AUC = 0.67 \pm 0.1$ ), the criterion E ( $AUC = 0.75 \pm 0.14$ ) and the criterion G ( $AUC = 0.70 \pm 0.1$ ) achieved comparable performances to the human expert. Criterion B (intrusion symptoms) and C (avoidance symptoms) are the least learned criteria, with respectively  $0.63 \pm 0.14$  and  $0.55 \pm 0.09$  AUC scores.

Model Interpretation, after selecting the best classifier for a classification task using the ROC-AUC score average over 100 random runs, we average the feature importance scores using the appropriate method for each kind of model, see Fig. 5. All the interpretations should be analyzed by taking into account the performances presented in Table 9 for each classifier. We found that the most important features, on average, for predicting:

Textual features	p value	SE (cohen)	Power
Lexical significant features			
Model_DEATH	5.10 <sup>-3</sup>	- 0.6	0.7
Model_PHYSICAL_SENSATIONS	5.10 <sup>-2</sup>	- 0.5	0.5
Liwc_death	5.10 <sup>-2</sup>	- 0.5	0.4
Speech disfluencies significant features			
Score_disfluencies (...)	5.10 <sup>-2</sup>	- 0.5	0.5

**Table 5.** Significant (Mann–Whitney U statistical test) associations between Language Features and intrusion symptoms (Criterion B).

Textual features	p value	SE (Cohen)	Power
Lexical significant features			
Model_DEATH	5.10 <sup>-3</sup>	- 0.4	0.7
Liwc_death	5.10 <sup>-3</sup>	- 0.4	0.6
Morphosyntactic significant features			
PRON	2.10 <sup>-2</sup>	- 0.4	0.6
PROP (Proper Name)	2.10 <sup>-2</sup>	0.5	0.7
Model_ON_someone	3.10 <sup>-3</sup>	- 0.5	0.8
First_personal_pronoun_plur	3.10 <sup>-2</sup>	- 0.4	0.6

**Table 6.** Significant associations between Language Features and criterion D symptoms (Mann–Whitney U statistical test).

Textual features	p value	SE (Cohen)	Power
Sentences_number	6.10 <sup>-3</sup>	- 0.6	0.7
Words_number	2.10 <sup>-3</sup>	- 0.6	0.8
Sentiment and emotional significant features			
LabMT	2.10 <sup>-3</sup>	0.6	0.7
Feel_positive	9.10 <sup>-4</sup>	0.7	0.9
Textblob_polarity	6.10 <sup>-4</sup>	0.8	0.9
Lexical significant features			
Model_DEATH	8.10 <sup>-4</sup>	- 0.7	0.8
Model_BODY	9.10 <sup>-3</sup>	- 0.6	0.7
Model_PHYSICAL_SENSATIONS	1.10 <sup>-3</sup>	- 0.6	0.7
Morphosyntactic significant features			
Verb_indicatif_future	3.10 <sup>-3</sup>	- 0.6	0.8
Significant syntactic features			
Passive_count_norm	3.10 <sup>-3</sup>	- 0.5	0.5
Speech disfluencies significant features			
Score_generical_connector_matches	7.10 <sup>-3</sup>	0.7	0.8
Readability significant features			
Token_ratio_score	1.10 <sup>-3</sup>	0.8	0.9
Noun_ratio_score	3.10 <sup>-4</sup>	0.9	0.9
Adverb_ratio_score	4.10 <sup>-3</sup>	0.6	0.7
Discourse coherence significant features			
Graph_average_shrotest_path_g0	3.10 <sup>-3</sup>	0.7	0.9
Graph_Parrallel_Edge (PE)	2.10 <sup>-3</sup>	- 0.7	0.8
Graph_L2	1.10 <sup>-3</sup>	- 0.4	0.5
Graph_degree_average	1.10 <sup>-3</sup>	- 0.75	0.9
Graph_degree_std	1.10 <sup>-3</sup>	- 0.7	0.8

**Table 7.** Significant associations between Language Features and Criterion E symptoms (Mann–Whitney U statistical test).

Textual features	p value	SE (Cohen)	Power
Words_number	1.10 <sup>-2</sup>	- 0.5	0.6
Sentiment and emotional significant features			
Polarimot_negative	3.10 <sup>-3</sup>	0.6	0.8
Liwc_positive_emotion	7.10 <sup>-3</sup>	0.6	0.8
LabMT	1.10 <sup>-2</sup>	0.7	0.9
Gobin_positive	1.10 <sup>-2</sup>	0.6	0.9
Lexical significant features			
Model_DEATH	3.10 <sup>-5</sup>	- 0.8	0.9
Model_BODY	5.10 <sup>-4</sup>	- 0.6	0.9
Model_PHYSICAL_SENSATIONS	2.10 <sup>-4</sup>	- 0.7	0.9
Liwc_death	2.10 <sup>-3</sup>	- 0.4	0.5
Morphosyntactic significant features			
First_personal_pronoun_sing	3.10 <sup>-2</sup>	- 0.4	0.5
GENERICAL_PRESENT	1.10 <sup>-2</sup>	0.65	0.9
Significant syntactic features			
Passive_count_norm	2.10 <sup>-2</sup>	- 0.5	0.7
Readability significant features			
Token_ratio_score	3.10 <sup>-2</sup>	0.5	0.8
Discourse coherence significant features			
Graph_PE	1.10 <sup>-2</sup>	- 0.4	0.6
Graph_degree_average	4.10 <sup>-2</sup>	- 0.4	0.6
Graph_L2	1.10 <sup>-2</sup>	- 0.3	0.3

**Table 8.** Significant associations between Language Features and Functional significance (Criterion G) Symptoms (Mann–Whitney U statistical test).

Task	Models	Roc-Auc score
Full or partial PTSD	LR	<b>0.69 ± 0.09</b>
	RF	0.66 ± 0.10
	EBM	0.61 ± 0.08
Criterion B (Intrusions)	LR	<b>0.63 ± 0.14</b>
	RF	0.55 ± 0.16
	EBM	0.54 ± 0.14
Criterion C (avoidance)	LR	0.54 ± 0.09
	RF	<b>0.55 ± 0.09</b>
	EBM	0.50 ± 0.09
Criterion D (Dissociation)	LR	<b>0.67 ± 0.1</b>
	RF	0.57 ± 0.1
	EBM	0.56 ± 0.1
Criterion E (Hyperarousal)	LR	<b>0.75 ± 0.14</b>
	RF	0.73 ± 0.12
	EBM	0.72 ± 0.14
Criterion G (Functional significance)	LR	<b>0.70 ± 0.1</b>
	RF	0.67 ± 0.1
	EBM	0.67 ± 0.1

**Table 9.** Performances average on 100 random runs on the test dataset for each classification task and each model.

- Probable full or partial PTSD, are the proportion of truncations in the text (“wor...”), the lexical field associated with death, physical sensation, and body, the verb conjugated to past participle, direct discourse, historical present, and L2 and average node degree for the graph features.

[FR] PAR0798 : Moi j'avais, j'avais qu'une peur, je voyais bien que j'étais debout et que j'étais pas **blesée mortellement**, hein, mais j'avais qu'une peur, c'est **d'être défigurée**, parce que je voyais que, enfin je sentais que j'étais blessée à la **joue**. Je lui demandais : « est-ce que, est-ce que, est-ce que mon **visage** est entier, est-ce que j'ai pas, est-ce que j'ai pas la moitié de la **mâchoire arrachée** quoi ».

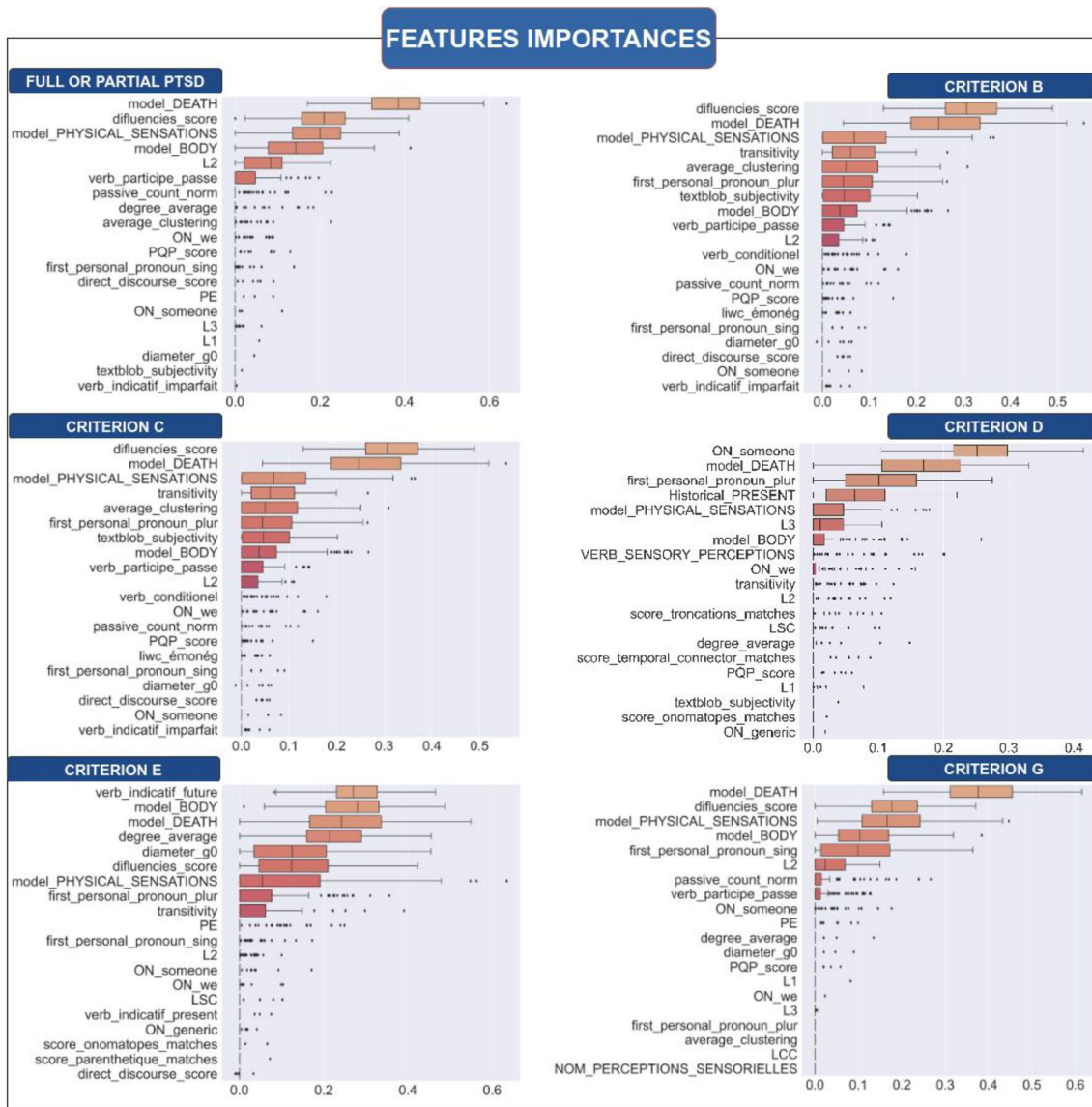
[EN] PAR0798: I had, I only had one fear, I could see that I was standing and that I wasn't **mortally wounded**, but I only had one fear, that I would be **disfigured**, because I could see that, well I felt that I had a wound on my **cheek**. I asked him: "Is my **face** whole, is it not, is half my **jaw torn off**?"

- Criterion B (intrusions), is, the proportion of disfluencies in the text (“wor...”), the lexical field associated with death, and physical sensation, the score of subjectivity from the textblob engine, and the transitivity and average clustering from the graph.

[FR] PAR0452: je suis sorti... ouais parce qu'il faut quand même que... donc lui, lui s'en est sorti aussi... indemne aussi... et en fait... oui ce qu'il se passe c'est qu'en fait, la première chose que je fais quand je rentre dans le taxi, une fois qu'on est... que le taxi se met en route...

[EN] PAR0452: I got out... yeah, because I have to... so he got out too... unharmed too... and in fact... yes, what happens is that, in fact, the first thing I do when I get into the taxi, once we're... once the taxi starts...

- Criterion C (avoidance), cannot be considered due to the performances' classifier that is near random.
- Criterion D (negative changes in cognition and mood), are lexical fields associated with death and physical sensations, verbs of perception, first person plural pronoun, and “ON\_someone”.



**Figure 5.** Features Importances of the diagnosis and all the criteria. The Feature importance is the average over 100 random runs for the model achieving the best results.

FR] PAR0618 : En attendant de nous transférer dans une caserne et après, à l'hôpital, on nous a mis dans une, une petite rue, une petite impasse (...) c'est là où on m'a fait un un bandage à la jambe.

EN] PAR0618: While waiting to transfer us to a barracks and afterwards, to the hospital, they put us in a, a small street, a small dead end (...) that's where they bandaged my leg.

- Criterion E (hyperarousal), are lexical fields of death, body and physical sensation, the future tense, the verbs of perceptions, and the average degree of nodes.

Pattern in French	Pattern in English
que ... C' est un peu de ma faute ... " Surtout que c' était ce était ... En plus ... j' ai oublié de parler	that ... It's a bit my fault... "Especially as it was ... Plus ... I forgot to speak
était étendu par terre , donc ... face contre terre et le visage ... en sang , enfin je pouvais ...	was lying on the floor, so... face down and... bleeding, well I could...
, mais , bon , ils s' occupaient des blessés et des morts , quoi , pas de	but, well, they were looking after the wounded and the dead, not the
... donc dans la fesse , et je me souviens , pareil , il réalise pas , je ... je	... so in the buttock, and I remember, same thing, he doesn't realize, I... I
le sang en fait ... d' Adrien ... moi je pensais qu' il allait mourir , parce que	the blood in fact ... of Adrien ... I thought he was going to die, because
le coup euh ... bah voilà Marie est décédée et puis ... dans ... c' était très fouillis , je me souviens	the blow ... well, Marie died and then ... in ... it was very messy, I remember
se passer , pourtant si ... euh ... Donc euh ... donc voilà on court	happen, although if ... er ... So um ... so here we are running
disant : " Non non c' était ... c' était rien , c' était ... des noeuds ou	saying: "No, no, it was ... it was nothing, it was ... knots or
due le coup ... et euh ... et je me souviens le ... lui dire : "	of the blow... and er... and I remember the... telling him: "
mon cerveau , il avait un peu , " clack " , il avait ... il avait débranché	my brain, it had a bit, "clack", it had ... it had disconnected

**Table 10.** Example of significant language pattern extracted by the CNN model.

[FR] MET0093: (...) je lui avais dit: "Tu sais (Nino) voilà s'il s'il arrive quelque chose, tu **jetteras** ton sac et on se **mettra** à courir très vite."

[EN] MET0093: (...) I said to him: "You know (Nino), if anything happens, you'll throw your bag away and we'll start running very fast."

- Criterion G (functional significance), is the lexical field of death, body and physical sensation, the first person singular pronoun, and the proportion of truncations in the narrative transcription.

[FR] Après coup, avec la peur, l'adrénaline, il avait **les jambes complètement tétanisées** (...) moi **je** me rappelle que j'étais j'étais très **oppressé**, **je** me rappelle j'avais comme un comme un espèce de de poids autour de la **poitrine**, j'avais du mal à bien faire rentrer l'air dans les **poumons**, j'avais du mal à bien inspirer quoi , c'est ça.

[EN] Afterwards, with the fear, the adrenaline, his **legs were completely tetanised** (...) I remember that **I** was very **oppressed**, **I** remember I had a sort of weight around my chest, I had difficulty getting the air into my **lungs**, I had difficulty breathing in properly, that's it.

### Error analysis

Statistical tests were carried out on the 20 most frequently misclassified documents over the 100 runs by the best classifier. The results show some statistical associations with the socio-demographic variables. The errors concerning the full or partial PTSD classifier are associated with the way of residence (p-value = 4.10<sup>-2</sup>, EF = 0.16, Power = 0.6). The classifier makes significantly fewer mistakes on people living alone. The errors for criterion B (intrusion symptoms) were associated with the type of exposition. We make significantly more mistakes in the A2 group (p-value = 4.10<sup>-5</sup>, EF = 0.3, Power = 0.98). The errors for criterion D (negative changes in cognition and mood) were associated with qualifications (p-value = 6.10<sup>-3</sup>, EF = 0.4, Power = 0.95) and matrimonial situation (p-value = 3.10<sup>-2</sup>, EF = 0.4, Power = 0.8). Indeed, the vast majority of mistakes concerned the "bac + 4 or more" subgroup and within the single population. The errors for criteria E (hyperarousal symptoms etc.) and G (functional significance etc.) were associated with the type of exposition. We make significantly more mistakes in the A2 group (p-value = 3.10<sup>-4</sup>, EF = 0.3, Power = 0.9) and respectively for the criterion G (functional significance etc.) (p-value = 3.10<sup>-3</sup>, EF = 0.2, Power = 0.9).

## Deep learning analysis

### Models performances

The CNN classifier performs worse than the psychiatrist or the machine learning classifier (see Supplementary Figs. S7 and S8). Particularly criteria C (avoidance), and D (negative changes in cognition and mood) are almost not learned on average and the performances on criteria B (Intrusion), E (hyperarousal), and G (functional significance) are nearly 0.1 down compared to other approaches. These underperformances might be explained by several reasons. Firstly, when constructing the dataset, the sequence breakdown can create false positives if the traumatic discourse is not uniformly distributed throughout the discourse. Secondly, previous approaches were guided by clinical knowledge that allowed us to construct relevant features and might explain the overperformances (Table 10).

Features group	Example of associated hypothesis	Hypothesis' origin	Extraction method
Sentiments and emotions	Traumatic language is associated with more negative emotions	Literature extracted	NLP literature resources
Lexical	Traumatic language is marked with some specific lexical fields such as death, body or physical sensations	Expert knowledge and literature extracted	Human annotation to train NER model
Morphosyntactic	Traumatic language is marked by a specific use of pronouns and verb tenses	Expert knowledge and literature extracted	Human annotation to train NER model and literature resources
Syntactic	Traumatic language is associated with an overuse of the passive voice	Expert knowledge	Literature resources
Speech disfluencies	The disfluencies are more frequent in traumatic language	Expert knowledge	Regular expressions
Readability	The retranscription of narratives from symptomatic people is less easy to read	Literature extracted	NLP literature resources
Narrative coherence	People with PTSD symptoms tends to produce less coherent narrative	Expert knowledge	Literature resources

**Table 11.** Description of the link between research hypothesis and Language Features.

### Model interpretations

Due to the lower performances of specific classifiers for each criterion, we focus on these parts on the interpretation of the main classifier that concerns full or partial PTSD. We selected a classifier that performs as well as the average performance and using the methodology proposed by Vanni et al.<sup>68–70</sup>, we extract the most representative sequence for the positive class Full or Partial PTSD and we can extract the most significant pattern using the Text Deconvolution Saillency applied on a multi-channel architecture<sup>69</sup>. The sequence of 512 tokens with the higher classification probability is presented here (see Supplementary Table S6 for French). While reading this sequence and the pattern in Table 11, we can observe that the model particularly focuses on disfluencies (euh, ...), repetitions, disorganized narrative, and onomatopoeia ('clack'). It also highlights direct discourse and some first names such as Adrien or Marie. Finally, we also observe some lexical fields that we expected such as death ('lying on the floor,' 'the dead'), and body parts ('buttock,' 'blood' and 'my brain'). We also note numerous figures of speech (metaphors: "it had disconnected.") and many words related to memory and memory impairments ("I remember"; it was very messy").

Most representative sequence translated into English: [...] a relationship with death, in those cases ... where well I'd seen others of the dead ... in the Bataclan ... and er ... of the good time OK ... Marie died, and it was more in relation to her family, where I was thinking about her family, saying to myself: "That's too stupid, she's made it this far and ... well, it's at the first-aid post where she died, even though it's the ... voilà. "So it's ... but I didn't ... I didn't think at the time that it could be me ... that, it was months later that I could think things like that, but ... not at the time. Er... at the time er... well, Marie died and then ... in ... it was very messy, I remember that the paramedics ... you could feel that there was a bit of panic, it's ... I saw them getting agitated but coming back to the same place fifteen times and poor Marie we came at least three, four times, to see if she was OK and then she was dead each time so um ... so PRON, I, I remember, also thinking to myself that, normally, we make colour codes for that, so that ... to identify the, the, the, the deceased, the ... the seriousness of the injured. So um ... so there you have it. And um ... and at the moment when I really began to find the time long, I ... I remember meeting the eye ... I perhaps stood up, a little ... and I met the eyes of a girl we were evacuating, and I remember that we looked at each other and that it was really ... a connivance, to think: "What the fuck are we getting out of here?" "To see ... yes, the look in our eyes as we couldn't talk to each other, because we were at the other end of the courtyard, from each other, but really from ... and then happy to catch someone's eye, too. So this girl left and then I was finally evacuated, but I was ... there weren't many people in the yard when I was evacuated ... So [...].

### Discussion

The blind evaluation provided evidence supporting the notion that language could serve as a marker of PTSD (ROC-AUC=0.72), while the ML method achieved comparable performance to the clinician's prediction (ROC-AUC=0.69), thus confirming and quantifying the language features associated with PTSD. The last study, utilizing the CNN, exhibited lower performance (ROC-AUC=0.64) but revealed language patterns that were not captured by the previous approaches. Regarding the underperformance of Random Forest and Explainable Boosting Machine, it aligns with existing literature. Typically, Random Forest and Explainable Boosting Machine outperform logistic regression (LR)<sup>34</sup>. However, in text classification benchmarks, LR tends to outperform other ML model<sup>71,72</sup>. It is noteworthy that although the ROC-AUC scores may not reach 1, they are comparable to the AUC values reported in the literature when comparing two Gold-Standard diagnostic methods for PTSD (PCL-5 and SCID). For example, Nedelcea et al.<sup>73</sup>, reported an average ROC-AUC of 0.76.

Supplementary Table S5 indicates that we have identified a range of markers representing a linguistic system associated with PTSD symptoms. Specifically, all examined dimensions of language are disrupted. These disruptions were related to specific PTSD symptoms. Language features, including lexical characteristics (e.g., lexicons

related to death, body, and physical sensations), emotional aspects (e.g., the emotional valence of discourse), morphosyntactic patterns (e.g., use of certain verb tenses and pronouns), syntactic aspects (e.g., the proportion of passive voice), and fluency-related elements (e.g., incomplete statements and repetitions), collectively contribute to identifying full or partial PTSD. Furthermore, discourse structures, as evidenced by graph-based features, are also altered in individuals with PTSD but further studies need to be carried out to provide more interpretable and theory-driven features for discourse structure.

Analyzing the sub-symptoms of PTSD based on the DSM-5 criteria, we found that reviviscence (Criterion B) is predominantly marked by lexical features, disfluencies, and perception-related verbs, which may indicate a literal re-experiencing of the traumatic scene. Negative changes in mood (Criterion D) is specifically identified by verbs of perception and the use of generic pronouns like "generic on" and "we," reflecting the depersonalization process. Hypervigilance, as represented by Criterion E, manifests through the use of perception-related verbs, future tense, historical present, and discourse that carries emotional intensity. The pronounced prevalence of the future tense among individuals experiencing symptoms of PTSD suggests that their anticipation of forthcoming activities diverges from that of psychologically sound populations, potentially stemming from involuntary future projections<sup>74</sup>. Impairment of global functioning (Criterion G) is linked to the frequent use of the first-person pronoun "I." This finding adds to the body of literature highlighting the excessive use of the first-person pronoun, a phenomenon commonly observed in individuals with depression, anxiety, and PTSD, which, in turn, is strongly associated with substantial clinical distress. It is worth mentioning that the difficulty in describing avoidance (Criterion C) may be correlated with this symptom and could reflect a bias in our dataset, as individuals participating in scientific studies may exhibit avoidance strategies.

Lexical features play a particularly discriminative role in the overall diagnosis of PTSD and each individual symptom. The mention of lexical fields related to death corresponds to Criterion A1 of PTSD, while the inclusion of lexical fields related to the "body" and "physical sensations" encompasses elements specific to the studied trauma (e.g., physical injuries, physical proximity in enclosed spaces) as well as physical manifestations of stress, anxiety, and negative emotions. We hypothesize that these lexical markers serve as reliable estimators of peritraumatic factors commonly assessed using tools like the Peritraumatic Distress Inventory<sup>75</sup>. One of the major peritraumatic risk factors associated with PTSD is loss of agentivity, which seems to be well captured by the proportion of passive forms in the narratives of our study. In contrast to the existing literature, we found no over-use of negative emotions. In our case, all the testimonials (PTSD or not) relate to a personal traumatic event, which could lead to homogeneity in the use of negative emotion terms. This is a major difference from the collections of tweets or reddit texts, in which other subjects were addressed by non-PTSD people<sup>19,34</sup>.

Additionally, graph-based features<sup>26</sup> effectively capture the structural characteristics of PTSD language. Indeed, measures such as L2, L3, and PE reflect narrative repetition and logorrhea, while average node degree, transitivity factor, and clustering coefficient reveal structural differences in narratives. In the forthcoming work, we will propose further exploration to better describe and understand how PTSD's discourse structure differs from that of non-PTSD.

The exploratory deep learning-based approach presented in this study uncovers the need for additional inquiry in unraveling the intricacies of traumatic language. Indeed, our qualitative analysis of the patterns extracted using TDS scores raises some considerations. First, there's the aspect of the low AUC score associated with the model. Second, there's the lack of quantitative analysis of the extracted patterns. Nevertheless, they provide avenues for further exploration. Specifically, we posit that figures of speech wield significant influence, yet their extraction using quantitative methodologies is challenged by their reliance on external world knowledge. Moreover, in alignment with the suggestions and practices of prior research<sup>75,76</sup>, we believe that a multimodal model incorporating speech and video data has immense potential in yielding insights in spite of drawbacks associated with data privacy.

Our research represents a notable advance in developing efficient language markers for PTSD. These markers provide low-cost, easily anonymized solutions with the potential to transform crisis response prioritization and patient monitoring. Their use will enable rapid identification of high-priority cases, thereby improving crisis intervention. They may also facilitate a thorough assessment of treatment effectiveness by tracking language markers during treatment, supporting more personalized patient care pathways and enhancing the objectivity of clinical judgment.

Methodologically, our study offers several contributions. First, we implemented a design that leverages the expertise of different professionals (psychiatrists and linguists) to produce scientific results. Second, by constructing highly discriminative and interpretable features using custom models based on expert knowledge, we demonstrated the potential of this approach for other pathologies, such as depression or psychosis.

## Limitations

Despite the meticulous application of our scientific methodology, certain limitations in our study could not be entirely mitigated. Firstly, the size of our dataset remains relatively small, which may restrict the generalizability of our findings regarding the association between language and PTSD, although it is one of the largest corpora in the existing literature. Additionally, the recruitment process employed might have introduced some bias into our sample, as individuals who experienced more intense trauma are often less inclined to participate in psychological trauma studies. We believe that a specific scientific design should be built to study precisely this dimension of avoidance. Particularly, the avoidance patterns might not have been well captured by textual data and could be better captured with video and audio data. This can also explain why clinical experts had difficulty in assessing this symptom using only transcriptions. Moreover, the study design did not encompass precise control over comorbidities and did not measure the effect of treatment, which could potentially influence the observed language patterns such as depression or anxiety. Moreover, as it is a time-consuming task, the blind

assessment was only carried out by one psychiatrist, which weakens the results. Lastly, the NLP methodology developed was focused on interpretation and transparency, and more recent methods such as Large Language Model or deeper neural model can improve classification performance (experiments are reported in the SI using Mistral-7B<sup>76</sup> and other architectures, Supplementary Tables E3, E4, E5).

## Perspective

Although our study represents a valuable advancement in the development of efficient language markers for PTSD, future studies could integrate multimodal data to emphasize the analysis of disfluencies and avoidance patterns. While our models show promise for diagnosing PTSD in survivors of terrorist attacks, further validation in diverse trauma populations is necessary to assess their broader applicability and to propose reliable clinical use. The first step would be to test the generalizability of our model to the rest of the Etude 1000 cohort (indirectly exposed individuals), and then to apply it to cohorts with different sources of PTSD (such as car accidents, war zones, or sexual violence). The next studies will also investigate how the linguistic markers can be used as predictive factors by exploring the second and third phases of the longitudinal Etude 1000, which took place in 2018 and 2021.

## Data availability

The participants of this study did not give written consent for their data to be shared publicly, due to the sensitive nature of the research, then, supporting data is not available. However, the anonymized and non-identifying text features are available and can be downloaded using the link in Supplementary Table S0 SM. The anonymized Text features dataset is available on this GitHub repository: [text\\_features.csv](#). The Anonymized psycho-socio-economic dataset: [socio\\_psycho\\_dataset.csv](#).

Received: 13 December 2023; Accepted: 7 May 2024

Published online: 30 May 2024

## References

- Auxéméry, Y. & Gayraud, F. Identification des marques du traumatisme psychique dans le langage parlé : définition de l'échelle diagnostique « SPLIT-10 ». *Ann. Méd.-Psychol. Rev. Psychiatr.* **179**, 869–888 (2021).
- Kleim, B., Horn, A. B., Kraehenmann, R., Mehl, M. R. & Ehlers, A. Early linguistic markers of trauma-specific processing predict post-trauma adjustment. *Front. Psychiatry* **9**, 400714 (2018).
- Bullock, A. *et al.* 1076 Self-referential language in trauma narratives predicts shorter sleep duration in women with Ptsd. *Sleep* **43**, A410 (2020).
- Zasiekina, L. Trauma, rememory and language in Holodomor survivors' narratives. *Psycholinguistics* **27**, 80–94 (2020).
- Pennebaker, J., Kiecolt-Glaser, J. & Glaser, R. Disclosure of traumas and immune function: Health implications for psychotherapy. *J. Consult. Clin. Psychol.* **56**, 239–245 (1988).
- Auxéméry, Y. Treatment of post-traumatic psychiatric disorders: A continuum of immediate, post-immediate and follow-up care mediated by specific psychotherapeutic principles. Clinical experience in French-speaking countries. *L'Encéphale* **44**, 403–408 (2018).
- Gayraud, F. & Auxéméry, Y. Identification of the marks of psychic trauma in spoken language: Definition of the "SPLIT-10" diagnostic scale. *Ann. Méd.-Psychol. Rev. Psychiatr.* **180**, 195–212 (2022).
- Breslau, N., Lucia, V. C. & Davis, G. C. Partial PTSD versus full PTSD: An empirical examination of associated impairment. *Psychol. Med.* **34**, 1205–1214 (2004).
- Zlotnick, C., Franklin, C. L. & Zimmerman, M. Does 'subthreshold' posttraumatic stress disorder have any clinical relevance?. *Compr. Psychiatry* **43**, 413–419 (2002).
- Mary, A. *et al.* Resilience after trauma: The role of memory suppression. *Science* **367**, eaay8477 (2020).
- Motreff, Y. *et al.* Factors associated with PTSD and partial PTSD among first responders following the Paris terror attacks in November 2015. *J. Psychiatr. Res.* **121**, 143–150 (2020).
- Sareen, J. Posttraumatic stress disorder in adults: Impact, comorbidity, risk factors, and treatment. *Can. J. Psychiatry Rev. Can. Psychiatr.* **59**, 460–467 (2014).
- Low, D. M., Bentley, K. & Ghosh, S. Automated assessment of psychiatric disorders using speech: A systematic review. <https://doi.org/10.31219/osf.io/5pwze> (2019).
- Argolo, F. *et al.* Lowering costs for large-scale screening in psychosis: A systematic review and meta-analysis of performance and value of information for speech-based psychiatric evaluation. *Braz. J. Psychiatry* **42**, 673–686 (2020).
- Le Glaz, A. *et al.* Machine learning and natural language processing in mental health: Systematic review. *J. Med. Internet Res.* **23**, e15708 (2021).
- Zhang, T., Schoene, A. M., Ji, S. & Ananiadou, S. Natural language processing applied to mental illness detection: A narrative review. *NPJ Digit. Med.* **5**, 1–13 (2022).
- Ramos-Lima, L. F., Waikamp, V., Antonelli-Salgado, T., Passos, I. C. & Freitas, L. H. M. The use of machine learning techniques in trauma-related disorders: A systematic review. *J. Psychiatr. Res.* **121**, 159–172 (2020).
- Alam, M. A. U. & Kapadia, D. LAXARY: A trustworthy explainable twitter analysis model for post-traumatic stress disorder assessment. *ArXiv200307433 Cs* (2020).
- Coppersmith, G., Harman, C. & Dredze, M. Measuring post traumatic stress disorder in Twitter. *ICWSM* (2014).
- Reece, A. G. *et al.* Forecasting the onset and course of mental illness with Twitter data.
- Jackson, R. G. *et al.* Natural language processing to extract symptoms of severe mental illness from clinical text: The Clinical Record Interactive Search Comprehensive Data Extraction (CRIS-CODE) project. *BMJ Open* **7**, e012012 (2017).
- He, Q., Veldkamp, B. P., Glas, C. A. W. & de Vries, T. Automated assessment of patients' self-narratives for posttraumatic stress disorder screening using natural language processing and text mining. *Assessment* **24**, 157–172 (2017).
- Calvo, R. A., Milne, D. N., Hussain, M. S. & Christensen, H. Natural language processing in mental health applications using non-clinical texts. *Nat. Lang. Eng.* **23**, 649–685 (2017).
- Son, Y. *et al.* World Trade Center responders in their own words: Predicting PTSD symptom trajectories with AI-based language analyses of interviews. *Psychol. Med.* <https://doi.org/10.1017/S0033291721002294> (2021).
- Blevins, C. A., Weathers, F. W., Davis, M. T., Witte, T. K. & Domino, J. L. The posttraumatic stress disorder checklist for DSM-5 (PCL-5): Development and initial psychometric evaluation. *J. Trauma. Stress* **28**, 489–498 (2015).
- Mota, N. B. *et al.* Speech graphs provide a quantitative measure of thought disorder in psychosis. *PLoS One* **7**, e34928 (2012).

27. Corcoran, C. M. *et al.* Prediction of psychosis across protocols and risk cohorts using automated language analysis. *World Psychiatry* **17**, 67–75 (2018).
28. Gupta, A., Kaur, M., Bajaj, A. & Khanna, A. Entailment and spectral clustering based single and multiple document summarization. *Int. J. Intell. Syst. Appl.* **11**, 39–51 (2019).
29. Pennebaker, J., Booth, R. & Francis, M. Linguistic inquiry and word count (LIWC2007). (2007).
30. Dodds, P. S., Harris, K. D., Kloumann, I. M., Bliss, C. A. & Danforth, C. M. Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. *PLoS One* **6**, e26752 (2011).
31. Akbik, A. *et al.* FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)* 54–59 (Association for Computational Linguistics, 2019). <https://doi.org/10.18653/v1/N19-4010>.
32. Schwartz, H. A. *et al.* DLATK: Differential language analysis ToolKit. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* 55–60 (Association for Computational Linguistics, 2017). <https://doi.org/10.18653/v1/D17-2010>.
33. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
34. Reece, A. G. *et al.* Forecasting the onset and course of mental illness with Twitter data. *Sci. Rep.* **7**, 13006 (2017).
35. Chandler, C., Foltz, P., Cohen, A., Holmlund, T. & Elvevåg, B. Safeguarding against spurious AI-based predictions: The case of automated verbal memory assessment. 181–191 (2021). <https://doi.org/10.18653/V1/2021.CLPSYCH-1.20>.
36. Chandler, C., Foltz, P. W. & Elvevåg, B. Using machine learning in psychiatry: The need to establish a framework that nurtures trustworthiness. *Schizophr. Bull.* **46**, 11–14 (2020).
37. da Rocha, D. *et al.* Programme 13-Novembre, un an après. *Rev. Neuropsychol.* **8**, 227–232 (2016).
38. Eustache, F. & Peschanski, D. Toward new memory sciences: The Programme 13-Novembre. *Prog. Brain Res.* **274**, 177–201 (2022).
39. American Psychiatric Association & American Psychiatric Association. DSM-5 Task Force. *Diagnostic and Statistical Manual of Mental Disorders: DSM-5*.
40. Greenhoot, A. F., Sun, S., Bunnell, S. L. & Lindboe, K. Making sense of traumatic memories: Memory qualities and psychological symptoms in emerging adults with and without abuse histories. *Memory* **21**, 125–142 (2013).
41. Piolat, A., Booth, R. J., Chung, C. K., Davids, M. & La Pennebaker, J. W. version française du dictionnaire pour le LIWC: modalités de construction et exemples d'utilisation. *Psychol. Fr.* **56**, 145–159 (2011).
42. Gobin, P., Camblats, A.-M., Faurous, W. & Mathey, S. Une base de l'émotionnalité (valence, arousal, catégories) de 1286 mots français selon l'âge (EMA). *Rev. Eur. Psychol. Appliquée/European Rev. Appl. Psychol.* **67**, 25–42 (2017).
43. Abdaoui, A., Azé, J., Bringay, S. & Poncelet, P. FEEL: A French expanded emotion lexicon. *Lang. Resour. Eval.* **51**, 833–855 (2017).
44. Fast, E., Chen, B. & Bernstein, M. Empath: Understanding topic signals in large-scale text. (2016). <https://doi.org/10.1145/2858036.2858535>.
45. Loria, S. textblob Documentation.
46. Martin, L. *et al.* CamemBERT: A Tasty French Language Model. <https://universalddependencies.org>.
47. Beaudreau, S. A. Are trauma narratives unique and do they predict psychological adjustment?. *J. Trauma. Stress* **20**, 353–357 (2007).
48. Qi, P., Zhang, Y., Zhang, Y., Bolton, J. & Manning, C. D. Stanza: A python natural language processing toolkit for many human languages. <https://doi.org/10.48550/arXiv.2003.07082> (2020).
49. Crespo, M. & Fernández-Lansac, V. Memory and narrative of traumatic events: A literature review. *Psychol. Trauma Theory Res. Pract. Policy* **8**, 149–156 (2016).
50. Jelinek, L. *et al.* Characteristics and organization of the worst moment of trauma memories in posttraumatic stress disorder. *Behav. Res. Ther.* **48**, 680–685 (2010).
51. Römisch, S., Leban, E., Habermas, T. & Döll-Hentschker, S. Evaluation, immersion, and fragmentation in narratives of distressing, angering, and happy events by traumatized and non-traumatized women. *Psychol. Trauma Theory Res. Pract. Policy* **6**, 465–472 (2014).
52. Arora, S., May, A., Zhang, J. & Ré, C. Contextual embeddings: When are they worth it? <https://doi.org/10.48550/arXiv.2005.09117> (2020).
53. Smirnova, D. *et al.* Language patterns discriminate mild depression from normal sadness and euthymic state. *Front. Psychiatry* **9**, 105 (2018).
54. Rabellino, D. *et al.* Altered sense of body ownership and agency in posttraumatic stress disorder and its dissociative subtype: A rubber hand illusion study. *Front. Hum. Neurosci.* **12**, 163 (2018).
55. Sepehri, A., Markowitz, D. M. & Mirshafiee, M. S. PassivePy: A tool to automatically identify passive voice in big text data. <https://doi.org/10.31234/osf.io/bwp3t> (2022).
56. de Boer, J. N., Brederoo, S. G., Voppel, A. E. & Sommer, I. E. C. Anomalies in language as a biomarker for schizophrenia. *Curr. Opin. Psychiatry* <https://doi.org/10.1097/YCO.0000000000000595> (2020).
57. Auxéméry, Y. & Gayraud, F. Le syndrome psycholinguistique traumatique (SPLIT). *L'Évolution Psychiatr.* **85**, 509–528 (2020).
58. Elvevåg, B., Foltz, P. W., Weinberger, D. R. & Goldberg, T. E. Quantifying incoherence in speech: An automated methodology and novel application to schizophrenia. *Schizophr. Res.* **93**, 304–316 (2007).
59. Bedi, G. *et al.* A window into the intoxicated mind? Speech as an index of psychoactive drug effects. *Neuropsychopharmacology* **39**, 2340–2348 (2014).
60. Hernandez, N., Faine, T. & Oulbaz, N. Open corpora and toolkit for assessing text readability in French. (2022).
61. McKnight, P. E. & Najab, J. Mann–Whitney U test. In *The Corsini Encyclopedia of Psychology* 1–1 (Wiley, 2010). <https://doi.org/10.1002/9780470479216.corpsy0524>.
62. Palmer, A. & Sesé, A. Recommendations for the use of statistics in Clinical and Health Psychology. *Clinica Salud* **24**, 47–54 (2013).
63. Nori, H., Jenkins, S., Koch, P. & Caruana, R. InterpretML: A unified framework for machine learning interpretability. <http://arxiv.org/abs/1909.09223> (2019).
64. Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **16**, 321–357 (2002).
65. Lin, T.-Y., Goyal, P., Girshick, R., He, K. & Dollár, P. Focal loss for dense object detection. <http://arxiv.org/abs/1708.02002> (2018).
66. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A. & Torralba, A. Learning deep features for discriminative localization. <http://arxiv.org/abs/1512.04150> (2015).
67. Zeiler, M. D. & Fergus, R. Visualizing and understanding convolutional networks. <http://arxiv.org/abs/1311.2901> (2013).
68. Vanni, L. *et al.* Text Deconvolution Saliency (TDS): A deep tool box for linguistic analysis. In *56th Annual Meeting of the Association for Computational Linguistics* (Melbourne, France, 2018).
69. Vanni, L., Corneli, M., Mayaffre, D. & Precioso, F. From text saliency to linguistic objects: Learning linguistic interpretable markers with a multi-channels convolutional architecture. *Corpus* <https://doi.org/10.4000/corpus.7667> (2023).
70. Vanni, L., Corneli, M., Longrée, D., Mayaffre, D. & Precioso, F. Key passages: From statistics to deep learning. In *Text Analytics. Advances and Challenges* (eds. Iezzi, D. F., *et al.*) 41–54 (Springer, 2020). [https://doi.org/10.1007/978-3-030-52680-1\\_4](https://doi.org/10.1007/978-3-030-52680-1_4).
71. Pranckevičius, T. & Marcinkevičius, V. Comparison of Naive Bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification. *Balt. J. Mod. Comput.* **5**, 221 (2017).
72. Shah, K., Patel, H., Sanghvi, D. & Shah, M. A comparative analysis of logistic regression, random forest and KNN models for the text classification. *Augment. Hum. Res.* **5**, 12 (2020).

73. Nedelcea, C. *et al.* The structure of PTSD. Development of the Post Traumatic Symptom scale from a clinician-based perspective. *Eur. J. Psychotraumatol.* **13**, (2022).
74. Finnbogadóttir, H. & Berntsen, D. Involuntary future projections are as frequent as involuntary memories, but more positive. *Conscious. Cogn.* **22**, 272–280 (2012).
75. Bunnell, B. E., Davidson, T. M. & Ruggiero, K. J. The Peritraumatic Distress Inventory (PDI): Factor structure and predictive validity in traumatically injured patients admitted through a Level I trauma center. *J. Anxiety Disord.* **55**, 8–13 (2018).
76. Jiang, A. Q. *et al.* *Mistral* 7B. <https://doi.org/10.48550/arXiv.2310.06825> (2023).

### Author contributions

R.Q. is the main author of this work, he developed the Python tools, wrote the Introduction, Methods, and Results Sections of the manuscript, and created all the figures and tables. F.G. contributed to the labeling task to train the token classifier, she was also deeply involved in the hypothesis generation process due to her experience in psycholinguistics. Y.A. as the main psychiatrist of this study realizes the blind evaluation process, writes most of the discussion, and contributes to making the manuscript more accessible to the clinician. L.V. as the main contributor of the TDS technics was in charge of the deep learning interpretation pipeline. All the authors approved the submitted version. D.P. et F.E. reviewed the manuscript and proposed significant changes in the form. J.D. as the second psychiatrist of the team was involved in all the discussions along the research process, he reviewed the manuscript and made significant changes to the Abstract and the Discussion to make it more impactful. He also contributed to labeling the diagnosis in the cohort based on the answers to the semi-standardized questionnaire. S.M. as the leader of the team was at the origin of the project, contributed to all discussions and deeply reviewed the manuscript, the code, and the analysis.

### Funding

The study was conducted within the framework of the “Programme 13-Novembre”, funded by the French General Secretariat for Investment (SGPI) through the National Research Agency (ANR) and the “Programme d’investissement pour l’Avenir” (PIA ANR-10-EQPX-0021-01). It also received the support of Region île-de-France through deliberation n°CP2021-297.

### Competing interests

The authors whose names are listed above certify that they have NO affiliations with or involvement in any organization or entity with any financial interest (such as honoraria; educational grants; participation in speakers’ bureaus; membership, employment, consultancies, stock ownership, or other equity interest; and expert testimony or patent-licensing arrangements), or non-financial interest (such as personal or professional relationships, affiliations, knowledge, or beliefs) in the subject matter or materials discussed in this manuscript.

### Additional information

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1038/s41598-024-61557-7>.

**Correspondence** and requests for materials should be addressed to R.Q.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher’s note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2024