Development and Validation of Deep Learning Transformer Models for Building a Comprehensive and Real-time Trauma Observatory

Gabrielle Chenais, Cédric Gil-Jardiné, Hélène Touchais, Marta Avalos Fernandez, Benjamin Contrand, Eric Tellier, Xavier Combes, Loick Bourdois, Philippe Revel, Emmanuel Lagarde

Development and Validation of Deep Learning Transformer Models for Building a Comprehensive and Real-Time Trauma Observatory

Gabrielle Chenais, Cédric Gil-Jardiné, Hélène Touchais, Marta Avalos Fernandez, Benjamin Contrand, Eric Tellier, Xavier Combes, Loick Bourdois, Philippe Revel, Emmanuel Lagarde

Submitted to: JMIR AI
on: July 07, 2022

Table of Contents

Development and Validation of Deep Learning Transformer Models for Building a Comprehensive and Real-Time Trauma Observatory

Gabrielle Chenais[1*] MMid, MScPH, MPHDs; Cédric Gil-Jardiné[1, 2*] MD, PhD; Hélène Touchais[1*] MCPM; Marta Avalos Fernandez[1, 3*] PhD; Benjamin Contrand[1*] MSc; Eric Tellier[2*] MD; Xavier Combes[2*] MD; Loick Bourdois[1*] MSc; Philippe Revel[2*] MD; Emmanuel Lagarde[1*] PhD

[1]Bordeaux Public Health Center, INSERM U1219 Bordeaux, FR
[2]Bordeaux University Hospital Emergency department Bordeaux FR
[3]Univeristy of Bordeaux, SISTM team, INRIA BSO Talence FR
[*]these authors contributed equally

Corresponding Author:
Gabrielle Chenais MMid, MScPH, MPHDs
Bordeaux Public Health Center, INSERM U1219
146 rue Léo Saignat
Bordeaux,
FR

Abstract

Background: In order to study the feasibility of setting up a national trauma observatory in France,

Objective: we compared the performance of several automatic language processing methods on a multi-class classification task of unstructured clinical notes.

Methods: A total of 69,110 free-text clinical notes related to visits to the emergency departments of the University Hospital of Bordeaux, France, between 2012 and 2019 were manually annotated. Among those clinical notes 22,481 were traumas. We trained 4 transformer models (deep learning models that encompass attention mechanism) and compared them with the TF-IDF (Term- Frequency - Inverse Document Frequency) associated with SVM (Support Vector Machine) method.

Results: The transformer models consistently performed better than TF-IDF/SVM. Among the transformers, the GPTanam model pre-trained with a French corpus with an additional auto-supervised learning step on 306,368 unlabeled clinical notes showed the best performance with a micro F1-score of 0.969.

Conclusions: The transformers proved efficient multi-class classification task on narrative and medical data. Further steps for improvement should focus on abbreviations expansion and multiple outputs multi-class classification.

(JMIR Preprints 07/07/2022:40843)
DOI: https://doi.org/10.2196/preprints.40843

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Original Paper

Gabrielle Chenais*, Univ. Bordeaux, INSERM, BPH U1219, F-33000 Bordeaux, France
gabrielle.chenais@u-bordeaux.fr

Cédric Gil-Jardiné, Bordeaux University Hospital, Emergency department, F-33000, Bordeaux, France
Univ. Bordeaux, INSERM, BPH U1219, F-33000 Bordeaux, France
cedric.gil-jardine@chu-bordeaux.fr

Hélène Touchais, Univ. Bordeaux, INSERM, BPH U1219, F-33000 Bordeaux, France
helene.touchais@u-bordeaux.fr

Marta Avalos Fernandez, Univ. Bordeaux, INSERM, BPH U1219, F-33000 Bordeaux, France
SISTM team, INRIA BSO, F-33405, Talence, France
marta.avalos-fernandez@u-bordeaux.fr

Benjamin Contrand, Univ. Bordeaux, INSERM, BPH U1219, F-33000 Bordeaux, France
benjamin.contrand@u-bordeaux.fr

Eric Tellier, Bordeaux University Hospital, Emergency department, F-33000, Bordeaux, France,
eric.tellier@chu-bordeaux.fr

Xavier Combes, Bordeaux University Hospital, Emergency department, F-33000, Bordeaux, France
xavier.combes@chu-bordeaux.fr

Loïck Bourdois, Univ. Bordeaux, INSERM, BPH U1219, F-33000 Bordeaux, France
loick.bourdois@hotmail.com

Philippe Revel, Bordeaux University Hospital, Emergency department, F-33000, Bordeaux, France
philippe.revel@chu-bordeaux.fr

Emmanuel Lagarde*
Univ. Bordeaux, INSERM, BPH U1219, F-33000 Bordeaux, France
emmanuel.lagarde@u-bordeaux.fr

Development and Validation of Deep Learning Transformer Models for Building a Comprehensive and Real-Time Trauma Observatory

Abstract

Background and Objective: In order to study the feasibility of setting up a national trauma observatory in France, we compared the performance of several automatic language processing methods on a multi-class classification task of unstructured clinical notes.

Methods: A total of 69,110 free-text clinical notes related to visits to the emergency departments of the University Hospital of Bordeaux, France, between 2012 and 2019 were manually annotated. Among those clinical notes 22,481 were traumas. We trained 4 transformer models (deep learning models that encompass attention mechanism) and compared them with the TF-IDF (Term-Frequency - Inverse Document Frequency) associated with SVM (Support Vector Machine) method.

Results: The transformer models consistently performed better than TF-IDF/SVM. Among the transformers, the GPTanam model pre-trained with a French corpus with an additional auto-supervised learning step on 306,368 unlabeled clinical notes showed the best performance with a micro F1-score of 0.969.

Conclusions: The transformers proved efficient multi-class classification on narrative and medical data. Further steps for improvement should focus on abbreviations expansion and multiple outputs multi-class classification.

Keywords: Deep Learning, Public Health, Trauma, Emergencies, Natural Language Processing, Transformers

Introduction

The objective of public health surveillance is to describe a health event in the population in order to estimate its burden based on its characteristics (incidence, prevalence, survival and mortality) and evolution. This surveillance contributes to the definition, implementation, monitoring and evaluation of public health policies. It must also be able to alert to the emergence of new threats to public health (infectious or environmental in origin, natural or terrorist) but also to monitor and evaluate the impact on the health of the population of known and expected events (seasonal epidemics) or unexpected events (industrial disasters, extreme weather events...). Public health surveillance relies on the collection of data, often in near-real time.

The SurSaUD (Surveillance Sanitaire des Urgences et des Décès) syndromic surveillance system was created for the purpose of public health surveillance in France in 2004 by Santé Publique France, the French National Public Health Agency. The SurSaUD system collects daily data from 4 data sources: Emergency Departments (OSCOUR ED network)[1], emergency general practitioners (SOS Médecins network), crude mortality (civil status data) and electronic death certification including causes of death.[2] Since its inception, the OSCOUR network has recorded more than 130 million ED visits. Data is collected by direct extraction of information from the patient's EHR (Electronic Health Record) in a common format for the whole territory and transmitted to Santé Publique France via the OSCOUR network. Thanks to the coding of the main diagnosis (ICD-10 codes - International

Classification of Diseases) and progressive improvement of data quality[3], the network can establish real-time monitoring of public health events such as epidemics of influenza, gastroenteritis or bronchiolitis.[4–7] This is one of the tools currently used to monitor responses to COVID-19 epidemic in France.

Approximately one-third of emergency department (ED) visits are the result of trauma in France.[8] Trauma is a major cause of mortality and morbidity in the world.[7] In 2017 in France, trauma and injury accounted for 7.01(6.75-7.33)% of deaths.[9] Unfortunately, little information is available when it comes to trauma: whereas we can know the nature of the main injury, nothing is known about the mechanism (road accident, assault, suicide ...). However, this information is available in the EHR, but in a free text form. In fact, each time a patient visits the ED, the nurse in charge of reception and orientation and the doctor in charge of the first consultation enter a text called clinical note, which describes the reasons for the patient's visit and the circumstances in which the symptoms occurred. In order to add the trauma mechanism to the data collected by the OSCOUR network, a manual classification by the health professionals would be time consuming and require multiple resources. Given the nature of the data (free text, unstructured and containing abbreviations) to be processed and the objective (classification), artificial intelligence with deep learning and more particularly automatic language processing seems to be indicated.

Natural language analysis has seen a recent breakthrough with the introduction of deep learning and in particular the transformer architecture. Introduced in 2017 by Google and proposed in the article Attention is All You Need by Vaswani et al.[10], transformers have an architecture that allows the implementation of a mechanism for processing the sequence of tokens[1] that form a sentence in a self-attentive way, i.e. relating each of these tokens to each of the others in the sentence. They have the particularity of being able to be pre-trained from a corpus of text which can be very large since it does not require a coding stage. This phase leads to a generative model which is capable, for example, of constructing artificial text by iteration. The Bidirectional Encoder Representations from Transformers (BERT) is one of these transformer-type models pre-trained on large corpora of text.[11] The BERT model is a bidirectional transformer, composed only of encoder blocks. Bidirectional indicates that BERT learns information from both the right and the left side of a token's context during the (pre)-training phase. BERT is composed of a stack of N =12 identical layers. Each layer has two sub-layers. The first is a multi-head self-attention mechanism, and the second is a simple, position-wise fully connected feed-forward network. In other words, text encoder turns text into a numeric representation. For many tasks, including text classification, its performance is systematically superior to the convolutional and auto-regressive models used until then.[11]

French derivatives such as FlauBERT[12] and CamemBERT[13] of the BERT model have been trained on very large and diverse French corpora. FlauBERT is a French BERT trained on a very large and heterogeneous French corpus. Models of different sizes were trained using the Jean Zay supercomputer of the CNRS (Centre national de la recherche scientifique). There are three sizes: small (54 million parameters), base cased (138M) and uncased (137M) as well as large (373M). CamemBERT is based on RoBERTa[14], an evolution of BERT in several aspects, including the use of the masked language model as the sole pre-training objective. CamemBERT, like FlauBERT, is available with different sizes: base (110M) and large (335M), but also with different training corpora such as OSCAR (either 138GB or 4GB of text)[15], CCNET (either 135GB or 4GB)[16] or French Wikipedia (4GB).

One of the most interesting examples of transformer architecture is GPT-2, released by OpenAI in 2019. GPT-2 (Generative Pre-Training 2) is a large transformer-based model, composed solely of decoder blocks, with 1.5 billion parameters on its extra-large version, trained on a dataset of 8 million web pages to predict the next word from the previous words.[17] Three other sizes of GPT-2 have been released before the largest one: with 124 (small), 355 (medium), 774 (large) million

[1] A token is an instance of a sequence of characters in a particular document that are grouped together as a semantic unit useful for language processing.

parameters. This model's ability for text generation quickly attracted the attention of the community because of the difficulty to distinguish the artificial texts produced from texts written by humans, suggesting that some of the meaning present in natural language was embedded. Moreover, beyond its ability to generate coherent texts, the GPT-2 can perform other tasks such as answering questions or classifying documents. As with BERT, the conservation of several self-attention blocks weights from a pre-trained model is sufficient to transfer contextual representations into another dataset. The training of the GPT-2 model is thus carried out in two distinct phases: the first phase of self-supervised generative pre-training, consists of the reading of a corpus of texts. It leads to the ability to generate texts automatically. The second supervised training phase consists in resuming the learning process from a corpus of annotated texts in order to create a system capable of performing specific tasks (classification for example). BelGPT2 is a Belgian small GPT2 pre-trained on French corpus of 60GB (Common Crawl, Project Gutenberg, Wikipedia, EuroPARL...) that was released at the end of 2020 .[18]

Related work

Extracting mechanisms and/or types of trauma is a matter of multi-class classification. Multi-class classification problems in French medical data have involved a wide variety of techniques. For example, for the 2018 CLEF eHealth Task 1 challenge[19] which objective is to extract ICD-10 codes from death certificates provided by the CépiDc (Centre for Epidemiology of Medical Causes of Death), the team of Cossin et al.[20] tested an approach based on ontologies, while Flicoteaux et al.[21] proposed an approach using a probabilistic CNN(Convolutional Neural Network) and Ive and al. [22]resorted to the association of a RNN(Recurrent Neural Network) with a CNN. On the other hand, Metzger et al. classified free-text clinical notes from ED relating to suicide attempts using Random Forest and Naive Bayes type algorithms.[23] Recent studies have shown the effectiveness of Transformers on classification tasks for EHR free-text data such as ICD coding[24,25], phenotyping[26] or readmission prediction[27]. Therefore, within the framework of the TARPON project (Traitement Automatique des Résumés de Passage aux urgences dans le but de créer un Observatoire National), which aims to demonstrate the feasibility of setting up a national observatory of trauma, we propose here to compare the performances of several transformer models for the classification of ED visits for trauma based on clinical notes from the adult emergency department of the Bordeaux University Hospital. We compared the transformers FlauBERT, CamemBERT, BelGPT2 and a French GPT2 model pre-trained on a domain-specific corpus called GPTanam here to the TF-IDF (Terms Frequency-Inverse Document Frequency)/SVM (Support Vector Machine) as a baseline model. To the best of our knowledge, no previous performance evaluation of multiple transformers for a classification application has been conducted on complex and unstructured clinical data from ED combining common French language, medical data and jargon.

Methods

Medical ethics regulations and GDPR

This study was authorized by the Bordeaux University Hospital Ethical Board under number GP-CE2021-21. A data management plan was created and reviewed by the privacy security board to meet institutional and national requirements in French for GDPR compliance.

Database

The clinical notes were extracted from the EHR of the adult emergency department stored in the

information system of the University Hospital of Bordeaux, France. They correspond to 375,478 medical records of visits to the adult emergency department of Bordeaux Hospital from 2012 to 2020. The variables available were age, sex, date and time of the visit, the clinical note generated by the doctors/interns and the clinical note written by the triage nurses.

Labeling strategy

69,110 clinical notes were randomly extracted for manual annotation. Our coding team consisted of trauma epidemiologists, emergency physicians, emergency nurses, research assistants, and biostatisticians, for a total of 16 coders. The annotation phase lasted 5 months. For each clinical note, a code describing the content of the text was assigned. The annotation grid used for the coding was developed for the needs of the project. The code associated with each clinical notes consisted of 9 fields. The fields were: "First visit (to the emergency department for this reason)", "Location (of the trauma)", "Activity (performed during the trauma)", "Type of Sport (practiced during the trauma)", "Subject under the influence", "Notion of pre-traumatic discomfort", "MVA (Motor Vehicle Accident)-Secondary Prevention Elements", "MVA-Antagonist", "Type of trauma or Mode of travel for the MVA". The objective being to classify the types of trauma, we used mainly the data of the field "Type of trauma or Mode of movement for the MVA". The distribution of the latter being unbalanced, we created a composite variable containing 8 mutually exclusive classes in order to have a larger number of clinical notes per class. Therefore, we grouped certain types of trauma (e.i. "Fall" which included "Fall from own height," "Fall from a given height," and "Fall on stairs"). The composite variable included the following classes/labels: "Accident of exposure to body fluids (blood exposure accident, unprotected sex at risk)" (AEF), "Assault", "Motor Vehicle Accident (MVA)", "Foreign body in eyes" (FBE), "Fall (except sports)", "Sports accident" (Acc. sport), "Intentional Injury", "Other trauma" as seen in Figure S1. The inter-annotator agreement was assessed with a random sample of 1000 clinical notes labelled by two annotators leading to a Cohen's kappa score[28] of 0.84.
A sensitivity analysis was performed in order to study the impact of potentially ambiguous content as regard to its classification. Therefore, the test sample was re-read by an expert. Potentially ambiguous content as regard to its classification is defined here as the accumulation of several mechanisms or types of trauma and/or a major difficulty in assigning a label to a clinical note given its text.

Corpus Statistics

In total, 22,481 manually labeled clinical notes from Bordeaux University Hospital were included in the study. Indeed, one-third (22,481/69,110) of the total annotated clinical notes were labeled as visit to the ED resulting from a trauma. The average number of sentences of the corpus was of 3.25 (min:1, max:63, std:2.56). The average length of clinical notes was of 58 words with a minimum of 1 (e.g., AES, Accident d'exposition au sang), a maximum of 630 and a standard deviation of 38 words. Unique unigrams, bigrams and trigrams were respectively equal to 70499, 395827 and 777459.

Models and experiment settings

The models selected for comparison and freely available as open-source content were a traditional machine learning (baseline model) with TF-IDF/SVM couple as well as 3 transformers type models pre-trained on French corpora: CamemBERT[13], FlauBERT[12]and BelGPT2[18]. We then chose the most performing model and applied a supplementary step of self-supervised training step with the

remaining 306,368 unlabeled clinical notes. This model is being called, here, GPTanam. Table 1 gives the size and configuration of each transformer model.

For the TF-IDF, tokenization was performed using the nltk v3.6.6 package[29] and linear SVC was applied with the use of scikit-learn v0.24.1[30]. The most frequent words (e.g., "that", "he", "the") were removed. Tokenization was performed with SentencePiece[31] for CamemBERT, Byte-Pair Encoding (BPE) for FlauBERT and a Byte-level BPE for both GPT-2 models[32]. The data was cleaned using regular expressions with the re package in Python 3.7. Unicode normalization was performed in UTF-8 (Universal Character Set Transformation Format - 8). Linear SVC parameters were as follows: tolerance = 1e-05, penalty = l2, loss = squared hinge, dual optimization = True, C=1.0, multi class strategy = one versus rest, verbose=0, a maximum of iterations of 1000. For all 3 transformers the optimizer was AdamW with an epsilon of 1e-8 and the maximum length was of 512. GPTanam had training and evaluation batch sizes of 5 and the learning rate was of 2e-5. For FlauBERT and CamemBERT, batch sizes for training was of 16 and 20 for evaluation and the learning rate was of 5e-5. Models were trained with the hugging face library under Pytorch framework on our workstation with a single Titan RTX (Nvidia©) GPU with 24GB of VRAM. Performance analysis was performed with scikit-learn and imbalance-learn v0.9.1

Table 1: Transformers models sizes and configurations. M: Millions, GB: Giga-Bytes

Model	Layers	Attention Heads	Embedding Dimension	Parameters	Pre-training Corpus size
CamemBERT-base-CCNET	12	12	768	110M	135GB
FlauBERT-base-cased	12	12	768	138M	71GB
BelGPT2	12	12	768	117M	57.9GB
GPTanam	12	12	768	117M	58.6GB

Table 2: Labels Distribution among Train, validation and test dataset

Type of Trauma	Train dataset		Validation dataset		Test dataset		Total	
	n	%	n	%	n	%	n	%
Accident of Exposure to Bodily Fluids	132	(0.9%)	40	(1.1%)	41	(1%)	213	(0.9%)
Assault	1587	(10.9%)	393	(10.8%)	498	(11.5%)	2478	(11%)
MVA	2028	(14%)	495	(13.6%)	568	(13.2%)	3091	(13.7%)

Foreign Body in Eye	642	(4.4%)	180	(5%)	186	(4.3%)	1008	(4.5%)
Fall	4778	(32.9%)	1162	(32%)	1554	(36%)	7494	(33.3%)
Sport Accident	1311	(9%)	341	(9.4%)	371	(8.6%)	2023	(9%)
Intentional Injury	341	(2.3%)	73	(2%)	112	(2.6%)	526	(2.3%)
Other trauma	3713	(25.6%)	950	(26.1%)	985	(22.8%)	5648	(25.1%)
Total	14532	(64.6%)	3634	(16.2%)	4315	(19.2%)	22481	(100%)

Self-supervised learning and Fine-tuning phase

Considering the GPTanam model, a first step comprising a self-supervised learning was performed with 306,368 clinical notes with one epoch.[33] For all models, a random sample of 80% (*n*= 18166) of the labeled as trauma (*n*=22481) was dedicated to supervised learning. This dataset was divided into a training sample (*n*=14532) and a validation sample (*n*=3634) with an 80/20 ratio. We trained each model 9 times with different seeds on 7 epochs for CamemBERT and FlauBERT and 5 epochs for BelGPT2 and GPTanam. In order to obtain a single prediction for the 9 different executions of the chosen epoch (based on maximum validation micro F1-score) for each model, a vote was taken.

Test phase

The test sample contained 20% of the labeled dataset, i.e. 4315 records. The second reading of these clinical notes resulted in 467 being tagged as clinical notes with potentially complex and/or ambiguous content as regard to its classification. The analysis therefore included both the complete test dataset (n=4315) and the dataset without complex and/or ambiguous content (n=3848). In order to obtain the probabilities for each prediction, a softmax activation layer was applied to the 4 transformer models.

Datasets

The label distribution among the corpus and each train, validation and test dataset is presented in Table 2. The most common type of trauma was the class "Fall" followed by "Other trauma" and "Motor Vehicle Accident". An example of clinical notes translated from French is given in supplementary Figure S2.

The median age at the visit was 37 years (1st and 3rd quartiles [24–58]) and 58.5% of patients were male. Electronic health record was introduced in year 2012 in Bordeaux University hospital, which explains the lower proportion of data for this particular year. Year 2019 saw a decrease in ED venues while in 2020 there have been a significant increase. Table 3 summarizes the characteristics of the train, validation and test datasets for the concerned population. Distribution of the variables age, sex and year of venues at the ED were comparable among the 3 datasets.

Table 3: Train, validation and test dataset characteristics

	Train dataset		Validation dataset		Test		Total	
Age	37	(24-58)	37	(24-57)	37	(24-58)	37	(24-58)
Sex. Male	8486	(58.3%)	2181	(59.9%)	2476	(57.5%)	13143	(58.5%)
Year of ED venue								
2012	218	(1.9%)	52	(1.8%)	66	(1.9%)	336	(1.9%)
2013	1389	(12.2%)	359	(12.4%)	418	(12.3%)	2166	(12.2%)
2014	1444	(12.6%)	385	(13.3%)	386	(11.3%)	2215	(12.3%)
2015	1502	(13.1%)	326	(11.2%)	425	(12.5%)	2253	(12.6%)
2016	1419	(12.4%)	365	(12.6%)	426	(12.6%)	2210	(12.3%)
2017	1493	(13.1%)	370	(12.8%)	461	(13.5%)	2324	(12.9%)
2018	1425	(12.5%)	405	(13.9%)	474	(13.9%)	2304	(13.5%)
2019	690	(6%)	175	(6%)	218	(6.4%)	1083	(6.2%)
2020	1856	(16.2%)	468	(16.1%)	532	(15.6%)	2856	(16%)
Missing values	3118	(27.3%)	737	(25.4%)	899	(26.4%)	4724	(20.9%)

Numbers are given along with percentages for sex and year of emergency department venue variables. Median, first quartile and fourth quartile are given for age.

Performance criteria

The measures chosen were macro-average precision and micro F1-score which, in the multi-class framework, is equal to accuracy. n: the number of samples (clinical notes), TP : True Positive, FP : False Positive, FN : False Negative.

Macro-average precision Precision expresses the proportion of units a model classifies as positive which are actually positive. In other words, precision tells us how much we can trust the model when it predicts a record is classified in a given class. In the case of multi-class classification, the macro-average precision over all classes i can be evaluated by macro-averaging, which first calculates the precision over each class i and then averages the precisions over all n classes. There is no relation to class size, as classes of different sizes are also weighted in the numerator. This implies that the effect of larger classes is as important as that of smaller classes. Each clinical note is therefore equally important with this measure.[34]

$$\text{MacroAverage } Precision = \frac{\sum_{i=1}^{n} precision_i}{n} \quad (1)$$

Micro F1-score F1-score is defined as a harmonic mean of precision and recall in binary class problem. To extend F1- measure to multi-class, two types of average, micro-average and macro-average are commonly used. In micro-averaging, the F1-measure is computed globally over all class decisions, precision and recall being obtained by summing over all individual decisions. Micro-averaged F1-measure gives equal weight to each clinical note and is therefore considered as an average over all the clinical note/category pairs.[35]

$$Micro\ F_1\ Score = \frac{2 * \sum_{i=1}^{n} TP_i}{2 * \sum_{i=1}^{n} TP_i + \sum_{i=1}^{n} FP_i + \sum_{i=1}^{n} FN_i} \setminus \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} TP_i + \sum_{i=1}^{n} FP_i} \quad (2)$$

Data availability

The dataset is not available due to patient privacy restrictions. However, the model may be shared with qualified researchers from academic or university institutions upon request via the corresponding author. Furthermore, identifying information was found in the dataset and a parallel work consisted in handling those using NER (Named Entity Recognition) and FlauBERT. Data processing and computing were conducted within the facilities of the Emergency Department of the University Hospital of Bordeaux which have received regulatory clearance to host and exploit databases with personal and medical data. All patients from which information were retrieved were 15 or more years old.

Error analysis

An error analysis was performed with uni and bigrams for the best performing model. All clinical notes misclassified were read by an expert to determine whether the human annotation label was appropriate or not.

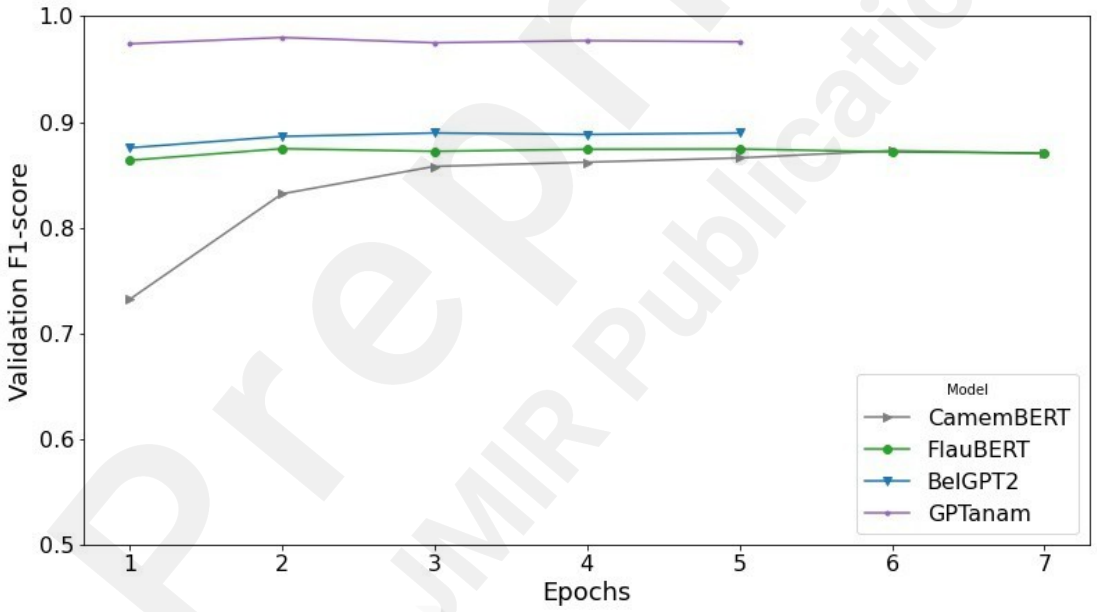
Results

Fine-tuning performance of models

Unlike machine learning methods such as TF-IDF, supervised fine-tuning of the transformer models is time-consuming and is greatly accelerated with the use of GPUs (Graphics Processing Unit). The self-supervised fine-tuning step for GPTanam model took approximately 12 hours. At that point, GPTanam could generate artificial clinical notes as seen in supplementary Figure S3, which could not easily be differentiated from the original ones. One epoch of supervised fine-tuning took 15, 16, 19 and 18 minutes for, respectively, CamemBERT, FlauBERT, BelGPT2 and GPTanam. When looking deeper into each transformer model's F1-scores on the validation dataset, Figure 1 shows that CamemBERT reached its maximum F1-score (0.873) at Epoch 6, FlauBERT achieved 0.874 at epoch 5, BelGPT2 was at its peak (0.890) faster at Epoch 3 and GPTanam reached 0.980 at epoch 2. Moreover, GPTanam's F1-score on the validation dataset was the highest among the 4 transformers models. We conjecture that self-supervised step on domain-specific corpus for GPTanam contributed

to a learning of the semantic representations which resulted in a faster convergence in the learning of the classification task.

Figure 1: F1-score curves for CamemBERT, FlauBERT, BelGPT2 and GPTanam on the validation dataset.



Performance of models

Average macro precision and micro F1-scores were systematically higher with the transformers than with the TF-IDF/SVM couple on the complete test dataset, as seen in Table. 4. Among the transformers, GPTanam achieved an average micro F1-score of 0.969, outperforming CamemBERT, FlauBERT and BelGPT2 for which F1-scores were 0.878, 0.873 and 0.887 respectively. Macro-average precision was higher than F1-score in almost all cases, except for TF-IDF/SVM where macro precision was lower than micro F1-score (macro precision = 0.860, micro F1-score = 0.864).
The distribution of *n* clinical notes per class not being balanced, the micro-F1 scores were, in all

cases, lower with the classes where *n* was lower. Concerning the micro F1-score of the different classes, GPTanam had higher scores than the other transformers and TF-IDF. The performance of GPTanam was high for all classes except for intentional injuries. We made the assumption that these results might be associated with the semantic heterogeneity and variety of this particular class. Indeed, this class encompassed self-arm (self-mutilation, punching due to rage, self-stabbing) and suicide attempts (shooting, alcohol or drug poisoning, car crashing) with few examples per injury. On the other hand, classes such as MVA or fall have semantic consistency with larger number of examples. The confusion matrix is given in Supplementary Figure S4. An the error analysis of the "intentional injury class, as well as the other classes, is provided in the next section.

Error analysis

Accident of exposure to bodily fluids: The unigram analysis showed that the key words "contact blood" were absent in the top 10 bigrams in the incorrectly classified clinical notes, while on the other hand unigrams analysis shows that "HIV" is the 9th unigram (after "aes", "blood", "needle", "source", "intercourse", "dakin", "work", "sexual").
Assault: Regarding the class "Assault", the top-3 bigrams were "physical assault", "declare having", and "punch" (fr:coup poing) for the correctly classified clinical notes while "left hand", "hand trauma", "mechanical fall" were the most frequent bigrams. The verification of the 18 clinical notes manually annotated as "Assault" showed that for 11 of them the label predicted by the model was correct (1 fall, 8 self-harm, 1 MVA, 1 sport accident paintball).
MVA: The acronym "mva" (n=700) was the most represented unigram in the correctly classified corpus while "pain" was the most represented one in the clinical notes classified as not MVA. When analyzing the 6 incorrectly classified clinical notes, 3 of them were wrongly labeled as they were in fact referring to an assault, a fall and a basketball accident. The 3 remaining clinical contained two types of trauma such as falling on the street.
Foreign body in the eye: The unigram analysis for this class showed that the unigrams "eye" and "theeye" were the most represented (n=140) while "left" and "hear" were the top-2 unigrams in the clinical notes classified as not being "foreign body in the eye". In fact, one of these clinical notes was related to a foreign body in the hear and two others were assault without mention of eye trauma.
Fall: The top-3 bigrams for the correctly classified clinical notes were "mechanical fall", "loss of consciousness", "cranial trauma" and were "right ankle", "ankle trauma", "left ankle" for the incorrectly classified ones. Twenty-one of the incorrectly classified clinical notes encompassed a double mechanism of trauma involving a sport accident, 16 MVA and 4 assault as well as a fall were present. Nine notes mentioned back pain, ankle and knee twists, pain while getting off of a truck, a patient found at the bottom of stairs, without mention of falling.
Intentional Injury: The most frequent uni and bigrams were different between the correctly and wrongly classified clinical notes. The most represented unigram and bigram were, respectively, "imv" (fr, voluntary drug intoxication) and "suicide attempt" in the correctly classified corpus of clinical notes while, "hand" and "punch given" were the most common in the correctly classified notes. Indeed, the model classified 10 clinical notes as assault while these clinical notes were related to a patient having punched something or himself.
Sport: The most frequent unigrams for correctly classified clinical notes were "pain", "left" and "trauma" and the bigrams were "right ankle", "functional impotence" and "left knee". The most frequent unigrams and bigrams for the incorrectly classified notes were, respectively "fall", "trauma", "bike" and "bike fall", "right knee", "knee pain". Thirteen falls occurred while biking without mention of the place and were classified as MVA. Five incorrectly classified notes were eye trauma while practicing sport.

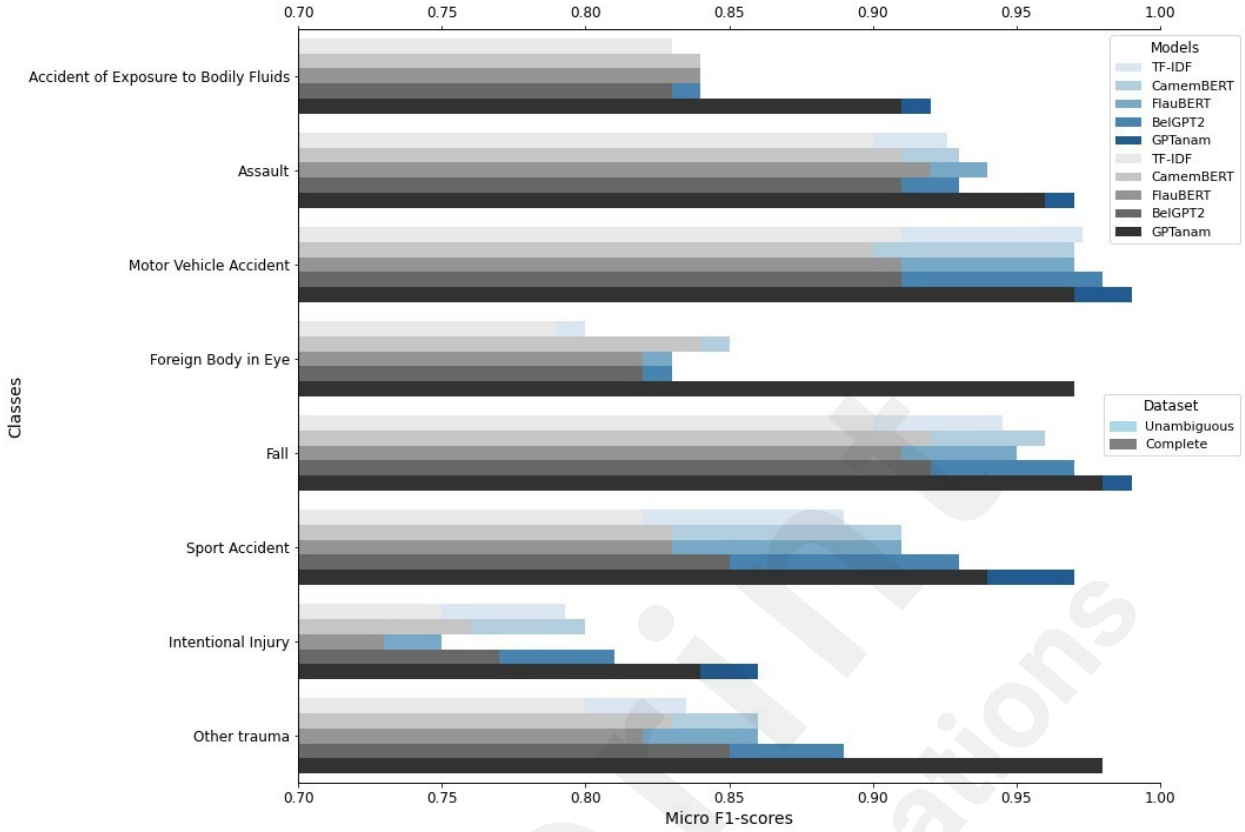
Table 4: Micro F1-scores for all classes and all models with micro average F1-scores and macro average precision on the complete test dataset.

Type of trauma	N	TF-IDF/ SVM	CamemBERT	FlauBERT	Bel-GPT2	GPTanam
Accident of Exposure to Bodily Fluids	41	0.83	0.84	0.84	0.83	**0.91**
Assault	498	0.9	0.91	0.92	0.91	**0.96**
MVA	568	0.91	0.90	0.91	0.91	**0.97**
Foreign Body in Eye	186	0.79	0.84	0.82	0.82	**0.97**
Fall	1554	0.9	0.92	0.91	0.92	**0.98**
Sport Accident	371	0.82	0.83	0.83	0.85	**0.94**
Intentional Injury	112	0.75	0.76	0.73	0.77	**0.84**
Other trauma	985	0.8	0.83	0.82	0.85	**0.98**
Total	4315					
Micro F1-score		0.864	0.878	0.873	0.887	**0.969**
Macro precision		0.860	0.880	0.880	0.89	**0.970**

Removing complex/ambiguous clinical notes is associated with an increase of performance for all models, the average gain of F1-scores being 0.04 for TF-IDF/SVM, CamemBERT, FlauBERT and BelGPT2. The average gain of micro F1-score was 0.01 for GPTanam, which seems more robust to complex and/or ambiguous content.

Difference in performance when potentially complex/ambiguous content is taken into account was greater with TF-IDF/SVM, CamemBERT, FlauBERT and BelGPT2 than with GPTanam, especially with the classes MVA and Sport Accident where the average gain of micro F1-score per class was 0.07 as seen in Figure 2. Performance for the class "Accident of exposure to bodily fluids" did not improve for TF-IDF/SVM, CamemBERT and FlauBERT when complex/ambiguous content was removed from the test dataset. Performance of GPTanam did not improve for the GPTanam with the classes "Foreign body on the eye" and "Other trauma" but F1-scores were already very high with, respectively, F1-scores of 0.97 and 0.98. Performance was slightly improved for 'Assault', 'Fall', 'MVA', 'Sport Accident' and 'Other trauma' when potentially complex and/or ambiguous content was removed from the test dataset for all models as seen in table Table S1 and confusion matrix in Figure S5.

Figure 2: Plot of micro F1-scores of all models for each class for both the complete test dataset (blue bars) and the test dataset without potentially ambiguous content as regard to its classification (grey bars).



Discussion

Transformers: a new state of the art

The transformers applied to the free text data from the ED of the Bordeaux University Hospital showed interesting results reaching an average micro F1-score of 0.969 for a GPT2 model with a French tokenizer and with a self-supervised training step on a domain specific corpus in addition to a large French corpus. This model showed better performance than TF-IDF/SVM and the other transformer models on average metrics and on all classes. In 2018, when reviewing deep learning algorithms for clinical NLP, Wu et al. projected the rise in popularity of Transformer models.[36] However, some studies show that traditional approaches, when tailored to the specific language and structure of the text inherent to the classification task, can achieve or exceed the performance of more recent ones based on contextual embeddings such as BERT. [37] Further study could involve comparing our model's performance to Bi-LSTM with pre-trained embeddings such as Word2Vec or transformers embeddings and CNN.

Self-supervised training on domain specific corpus and tokenizer

The decision to use pre-trained models on French corpora with a French tokenizer has probably contributed to the global performance of the chosen transformer models. General language transformer models pre-trained on a cross-domain text corpus in a given language have flourished recently. BelGPT2 was the first GPT2 model fine-tuned on a French heterogeneous corpus (CommonCrawl, French Wikipedia, EuroParl...) released on the hugging face platform. Self-supervised training of transformers on a specific domain can improve task performance such as

classification[38], text generation[39] and predicting hospital readmission.[40] Despite lots of experiments using BERT, GPT-2 hasn't been studies as well as BERT yet. Our team showed that the number of data required to achieve a given level of performance (area under the curve over 0.95) was reduced by a factor of 10 when applying self-supervised training on emergency clinical notes to a binary classification task.[41] Here, we confirm the benefits of a self-supervised training step on a domain-specific corpus. However, it is questionable whether this approach will be applicable when extending the TARPON project to data from other EDs in France, as each region or ED uses a specific language in addition to the medical language, which uses many abbreviations that can vary locally (i.e. assault is written "brawl" in Bordeaux, "hep" means hepatitis...). A possible solution would be to train the model on a corpus resulting from the extraction of ED notes at a national level. Similarly, the treatment of medical concepts and abbreviations remains an area for improvement, as not all EDs use the same abbreviations in the same context. The use of ontologies developed in the field of emergencies could constitute an area for improvement. Transformers have also recently been tested for the identification and replacement of abbreviations with good results for BERT[42,43], however there has not yet been a test on data from a mixture of common language and medical terms in French.
In addition, as the authors who proposed the CamemBERT model did not compare the different models from the OSCAR, CCNet and Wikipedia datasets on a classification task, a future work could compare the different sets on our database. In this logic, it would be appropriate, while we have only used the basic models of CamemBERT, FlauBERT and GPT-2, to test the different sizes of pre-training datasets on a classification task as well as the different sizes of models. Indeed, Martin's team has shown that the standard CamemBERT model (110 million parameters) trained on all 138GB of OSCAR text, does not massively outperform the model trained "only" on the 4GB sample in morphosyntactic labeling, syntactic parsing, Named Entity Recognition (NER) and Natural Language Inference (NLI).[44] One perspective considered is to test different models of French transformers that have been released since CamemBERT, FlauBERT or BelGPT2 such as Pagnol or BARThez.

Taxonomy

The performance of the models improved when we excluded the clinical notes that we considered to be the most complex and/or ambiguous from our test dataset. The classification errors analysis showed that when clinical notes encompassing two mechanisms of trauma (e.i. "fall from bike on the street") were removed from the test dataset, models performed better. This expected result shows that since the advent of transformers, the margin of progress in a free text classification task is nowadays low. This behavior was less important with GPTanam, which seems to have benefited from the self-supervised pre-training phase for reducing classification errors by learning semantic representations beforehand. However, the annotation grid created for the project is partly responsible for some classification errors in the sense that there are areas of semantic overlap between classes. In addition, the coding system used did not allow for the coding of several traumatic mechanisms (e.g., a collision between two individuals, followed by a fall). In order to be able to account for these situations, a new coding system will be used for the next phases of the project, using the recently released version of trauma classification grid used by the FEDORU (Fédération des Observatoires Régionaux des Urgences) and OSCOUR.

Improving Trauma Public Health Surveillance

The costs of injury and morbidity are immense, not only in terms of lost economic opportunities and demands on national health budgets, but also in terms of personal suffering.[45] However, few

countries have surveillance systems that generate reliable information on the nature and extent of injuries, especially with regards to non-fatal injuries. The traditional view of injuries as “accidents” or random events has resulted in the historical neglect of this area of public health.[46] Yet, for the last decades, public health officials have been recognizing traumas as preventable events and have been promoting evidence-based interventions for the prevention of traumas worldwide.[47] Many injury interventions are already in place (e.g., transportation requirements such as setting speed limits, safe automobile design, seatbelt and other safety restraint use, helmet and protective equipment) and achieved significant public health improvements including reduction of trauma occurrence.[48]
The automatic labelling of ED clinical notes will contribute to an effective real-time public health surveillance system for traumas. Future steps encompass deployment in hospitals’ IT departments of Gironde, France at first, then at a national scale.

Conclusion

Transformers have shown great effectiveness in a multi-class classification task on complex data encompassing narrative, medical data and jargon. The choice of this type of architecture in the automatic processing of emergency department summaries in order to create a national observatory is relevant. Applying a self-supervised training step on a specific domain corpus has substantially improved classification performances with a French GPT2 model. The next labeling strategy within the framework of the TARPON project will be carried out using a standardized trauma classification tool, which will allow a more precise classification of trauma mechanisms due to a clearer delineation between the different classes (little overlap of semantic fields). The objective is eventually to have a single code for ED summaries including several variables (e.g., place of occurrence, activity during the trauma, role in a road accident). It will be necessary to investigate the possibility of making predictions with a model trained on each variable, or using a single model trained on all variables. If the latter method is chosen, a larger model of GPT2 will probably be required. Furthermore, the expansion of acronyms is under consideration in the automation pipeline.

AUTHORS CONTRIBUTIONS

E.L and G.C designed the experiments. G.C drafted the paper. H.T and G.C programmed the design and experiments. The scripts were checked together by H.T and G.C. G.C designed the dataset. C.G-J extracted the dataset from database. The paper was revised by all the authors. Guarantor is G.C.

Acknowledgment

This work was carried out within the framework of the TARPON project (Traitement Automatique des Résumés de Passages aux urgences pour un Observatoire National) led by the Inserm team Injury epidemiology (project leader E. Lagarde) and the emergency department of the Bordeaux University Hospital in collaboration with the SISTM team, shared by Inria and Inserm. This project is the winner of the 2nd call for projects of the Health Data Hub, Grand Défi “Improving medical diagnosis through Artificial Intelligence” and Bpifrance. This study has been carried within the framework of PIA3 (Investment for the Future), (project No. 17-EURE-0019). We would like to thank all members of the labeling team. We thank the University Hospital of Bordeaux for providing the logistical support that allowed us to access and analyse the data needed for the manuscript in such a short period. We are also grateful to Julien Anjoubault, Clarisse Marguinaud, Virginie Cocuelle, Delphine Vauthier, Alexandra Barbe, François Garreau, Quentin Bana, Claire Riou, Pauline Soubelet and Elisabeth Verbitskaya for their expertise, which allowed proper manual coding for validation and to Benjamin Contrand and Marie-Odile Coste for data management and administrative assistance. BPH IETO Team activities are supported by the Institut National de la Santé et de la Recherche Médicale

(INSERM), University of Bordeaux, Ministère de l'Intérieur (Délégation à la Sécurité Routière).

Abbreviations

BERT : Bidirectional Encoder Representations Transformer
CNN : Convolutional Neural Network
CNRS : Centre National de la Recherche Scientifique
GPT2 : Generative Pre-Trained Transformer 2
ICD : international classification of diseases
RNN : Recurrent Neural Network
SVM : support vector machine
TARPON : traitement automatique des résumés de passage aux urgences dans le but de créer un observatoire national
TF-IDF : term frequency-inverse document frequency

REFERENCES

1. Fouillet A, Fournet N, Caillère N, et al. *SurSaUD® Software: A Tool to Support the Data Management, the Analysis and the Dissemination of Results from the French Syndromic Surveillance System*.; 2012. http://ojphi.org
2. CASERIO SCHONEMANN C, HENRY V, FOUILLET A, BOUSQUET V. *Le Système de Surveillance Syndromique SurSaUDz*.; 2014. http://pascal-francis.inist.fr/vibad/index.php?action=getRecordDetail&idt=28195403
3. Josseran L, Fouillet A, Caillère N, et al. Assessment of a syndromic surveillance system based on morbidity data: Results from the Oscour® network during a heat wave. *PLoS One*. 2010;5(8). doi:10.1371/journal.pone.0011984
4. Paireau J, Pelat C, Caserio-Schönemann C, et al. Mapping influenza activity in emergency departments in France using Bayesian model-based geostatistics. *Influenza Other Respir Viruses*. 2018;12(6):772-779. doi:10.1111/irv.12599
5. Hughes HE, Morbey R, Fouillet A, et al. Retrospective observational study of emergency department syndromic surveillance data during air pollution episodes across London and Paris in 2014. *BMJ Open*. 2018;8(4). doi:10.1136/bmjopen-2017-018732
6. Subiros M, Brottet E, Solet JL, Leguen A, Filleul L. Health monitoring during water scarcity in Mayotte, France, 2017. *BMC Public Health*. 2019;19(1). doi:10.1186/s12889-019-6613-8
7. *Global, Regional, and National Disability-Adjusted Life-Years (DALYs) for 359 Diseases and Injuries and Healthy Life Expectancy (HALE) for 195 Countries and Territories, 1990-2017: A Systematic Analysis for the Global Burden of Disease Study 2017*.; 2018. https://vizhub.healthdata.org
8. Cour des comptes. *Annual Public Report 2019: Hospital Emergency Departments*.; 2019. Accessed March 3, 2020. www.ccomptes.fr.
9. Global Burden of Disease Project. Accessed March 1, 2020. http://www.healthdata.org/gbd
10. Vaswani A, Brain G, Shazeer N, et al. Attention Is All You Need. In: *Advances in Neural Information Processing Systems*. ; 2017. doi:10.48550/arXiv.1706.03762
11. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Published online October 10, 2018. http://arxiv.org/abs/1810.04805
12. Le H, Vial L, Frej J, et al. FlauBERT: Unsupervised Language Model Pre-training for French. Published online December 11, 2019. http://arxiv.org/abs/1912.05372
13. Martin L, Muller B, Suárez PJO, et al. CamemBERT: a Tasty French Language Model. Published online November 10, 2019. doi:10.18653/v1/2020.acl-main.645

14. Liu Y, Ott M, Goyal N, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. Published online July 26, 2019. http://arxiv.org/abs/1907.11692
15. Javier Ortiz Suárez P, Sagot B, Romary L, Sagot B. Asynchronous Pipeline for Processing Huge Corpora on Medium to Low Resource Infrastructures. Published online 2019. doi:10.14618/IDS-PUB
16. Wenzek G, Lachaux MA, Conneau A, et al. CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data. Published online November 1, 2019. http://arxiv.org/abs/1911.00359
17. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language Models are Unsupervised Multitask Learners. *OpenAI blog*. 2019;1(8)(9). https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf
18. Louis A. BelGPT-2: a GPT-2 model pre-trained on French corpora. Published 2020. https://github.com/antoiloui/belgpt2
19. Suominen H, Kelly L, Goeuriot L, et al. Overview of the CLEF ehealth evaluation lab 2018. In: *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol 11018 LNCS. Springer Verlag; 2018:286-301. doi:10.1007/978-3-319-98932-7_26
20. Cossin S, Jouhet V, Mougin F, Diallo G, Thiessard F. IAM at CLEF eHealth 2018: Concept Annotation and Coding in French Death Certificates. Published online July 10, 2018. http://arxiv.org/abs/1807.03674
21. Flicoteaux R. ECSTRA-APHP @ CLEF eHealth2018-task 1: ICD10 Code Extraction from Death Certificates. In: *In CLEF (Working Notes).* ; 2018. http://www.cepidc.inserm.fr/
22. Amin-Nejad A, Ive J, Velupillai S. *Exploring Transformer Text Generation for Medical Dataset Augmentation*.; 2020. https://github.com/tensorflow/tensor2tensor
23. Metzger MH, Tvardik N, Gicquel Q, Bouvry C, Poulet E, Potinet-Pagliaroli V. Use of emergency department electronic medical records for automated epidemiological surveillance of suicide attempts: a French pilot study. *Int J Methods Psychiatr Res*. 2017;26(2). doi:10.1002/mpr.1522
24. Lopez-Garcia G, Jerez JM, Ribelles N, Alba E, Veredas FJ. Transformers for Clinical Coding in Spanish. *IEEE Access*. 2021;9:72387-72397. doi:10.1109/ACCESS.2021.3080085
25. Zhang Z, Liu J, Razavian N. BERT-XML: Large Scale Automated ICD Coding Using BERT Pretraining. Published online May 26, 2020. http://arxiv.org/abs/2006.03685
26. Liu Z, He H, Yan S, Wang Y, Yang T, Li GZ. End-to-end models to imitate traditional chinese medicine syndrome differentiation in lung cancer diagnosis: Model development and validation. *JMIR Med Inform*. 2020;8(6). doi:10.2196/17821
27. Mohammadi R, Jain S, Namin AT, et al. Predicting unplanned readmissions following a hip or knee arthroplasty: retrospective observational study. *JMIR Med Inform*. 2020;8(11). doi:10.2196/19761
28. Landis JR, Koch GG. *The Measurement of Observer Agreement for Categorical Data*. Vol 33.; 1977.
29. Bird S. NLTK: The Natural Language Toolkit. In: *Proceediproceedings of the COLING/ACL 2006 Interactive Presentation Sessions*. ; 2006:69-72. https://aclanthology.org/P06-4018.pdf
30. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*. 2011;12:2825-2830.
31. Kudo T, Richardson J. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. Published online August 19, 2018. http://arxiv.org/abs/1808.06226
32. Shibata Y, Kida T, Fukamachi S, et al. *Byte Pair Encoding: A Text Compression Scheme That Accelerates Pattern Matching*.; 1999. doi:10.1.1.46.4046
33. Komatsuzaki A. One Epoch Is All You Need. Published online June 16, 2019. http://arxiv.org/abs/1906.06669
34. Grandini M, Bagli E, Visani G. Metrics for Multi-Class Classification: an Overview. Published online August 13, 2020. http://arxiv.org/abs/2008.05756

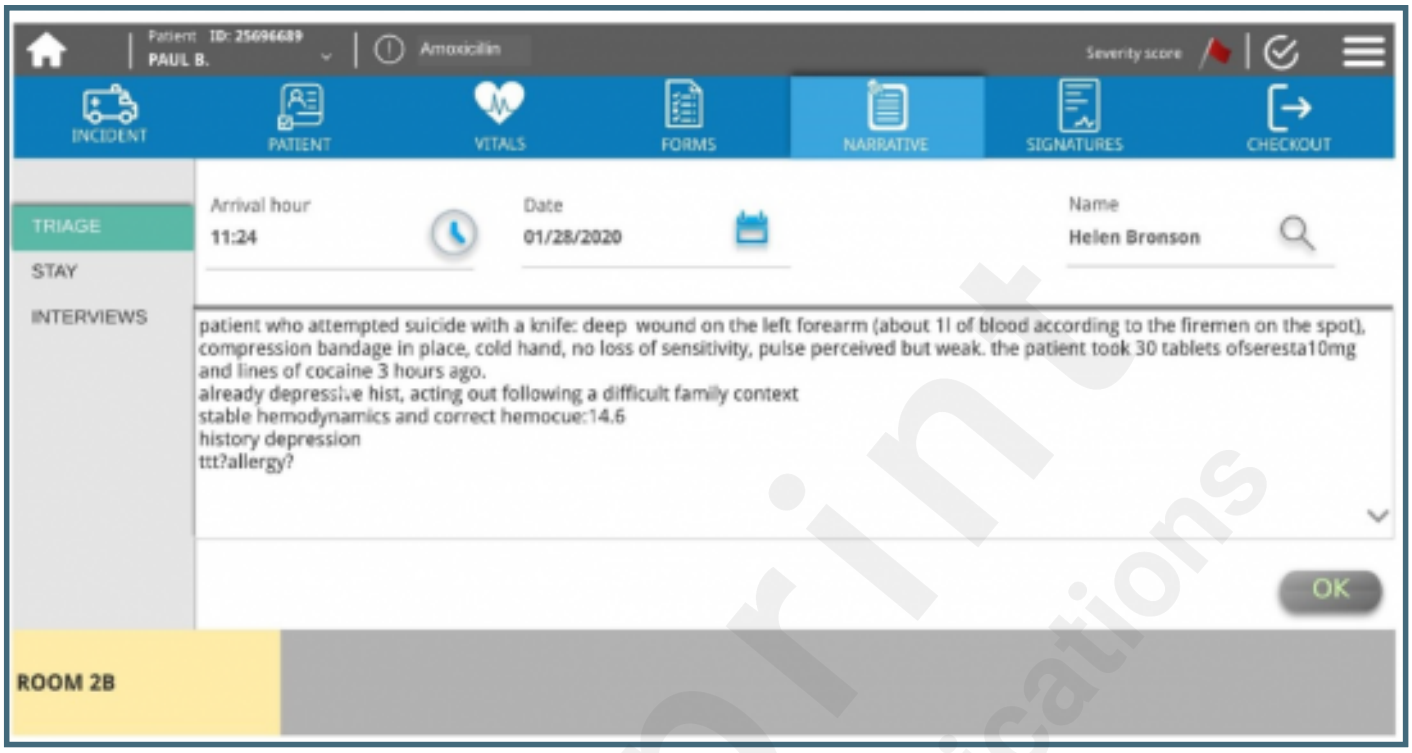
35. Arzucan¨ozgür, Levent¨ozgür L, Güngör T. Text Categorization with Class-Based and Corpus-Based Keyword Selection. In: *Computer and Information Sciences (ISCIS 2005)*. ; 2005:606-615.
36. Wu S, Roberts K, Datta S, et al. Deep learning in clinical natural language processing: A methodical review. *Journal of the American Medical Informatics Association*. 2020;27(3):457-470. doi:10.1093/jamia/ocz200
37. Mascio A, Kraljevic Z, Bean D, et al. Comparative Analysis of Text Classification Approaches in Electronic Health Records. Published online May 8, 2020. http://arxiv.org/abs/2005.06624
38. Yao L, Jin Z, Mao C, Zhang Y, Luo Y. Traditional Chinese medicine clinical records classification with BERT and domain specific corpora. *Journal of the American Medical Informatics Association*. 2019;26(12):1632-1636. doi:10.1093/jamia/ocz164
39. Lee JS, Hsiang J. Patent claim generation by fine-tuning OpenAI GPT-2. *World Patent Information*. 2020;62. doi:10.1016/j.wpi.2020.101983
40. Huang K, Altosaar J, Ranganath R. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. Published online April 10, 2019. http://arxiv.org/abs/1904.05342
41. Xu B, Gil-Jardine C, Thiessard F, Tellier E, Avalos-Fernandez M, Lagarde E. Pre-Training a Neural Language Model Improves the Sample Efficiency of an Emergency Room Classification Model. In: *The Thirty-Third International FLAIRS Conference (FLAIRS-33)*. ; 2020. www.aaai.org
42. Adams G, Ketenci M, Bhave S, et al. *Zero-Shot Clinical Acronym Expansion via Latent Meaning Cells*. Vol 136.; 2020. https://github.com/griff4692/LMC
43. Egan N, Bohannon J. Primer AI's Systems for Acronym Identification and Disambiguation. Published online December 14, 2020. http://arxiv.org/abs/2012.08013
44. Martin L, Muller B, Javier Ortiz Suárez P, et al. Les modèles de langue contextuels Camembert pour le français : impact de la taille et de l'hétérogénéité des données d'entrainement. Published online 2020. https://hal.archives-ouvertes.fr/hal-02784755v3
45. *WHO GUIDE TO IDENTIFYING THE ECONOMIC CONSEQUENCES OF DISEASE AND INJURY Department of Health Systems Financing Health Systems and Services*.; 2009.
46. Krug EG, Sharma GK, Lozano R. The global burden of injuries. *Am J Public Health*. 2000;90(4):523.
47. Organization WH. Injury surveillance guidelines / edited by: Y. Holder ... [et al.]. Published online 2001:Re-issued as a WHO publication.
48. Peden MM, World Health Organization., World Bank. *World Report on Road Traffic Injury Prevention*. World Health Organization; 2004.

Supplementary Files

Untitled.
URL: http://asset.jmir.pub/assets/16bdc0df85b068f191a9e0da6129c85c.doc

Figures

Supplementary Figure 2. Emergency Room Electronic Health Record visualization with clinical note translated in English.



Supplementary Figure 3. Example of 2 clinical notes artificially generated by GPTanam right after the self-supervised training step with a setup of maximum 40 tokens generated. Clinical notes in French on the left, translated in English on the right.

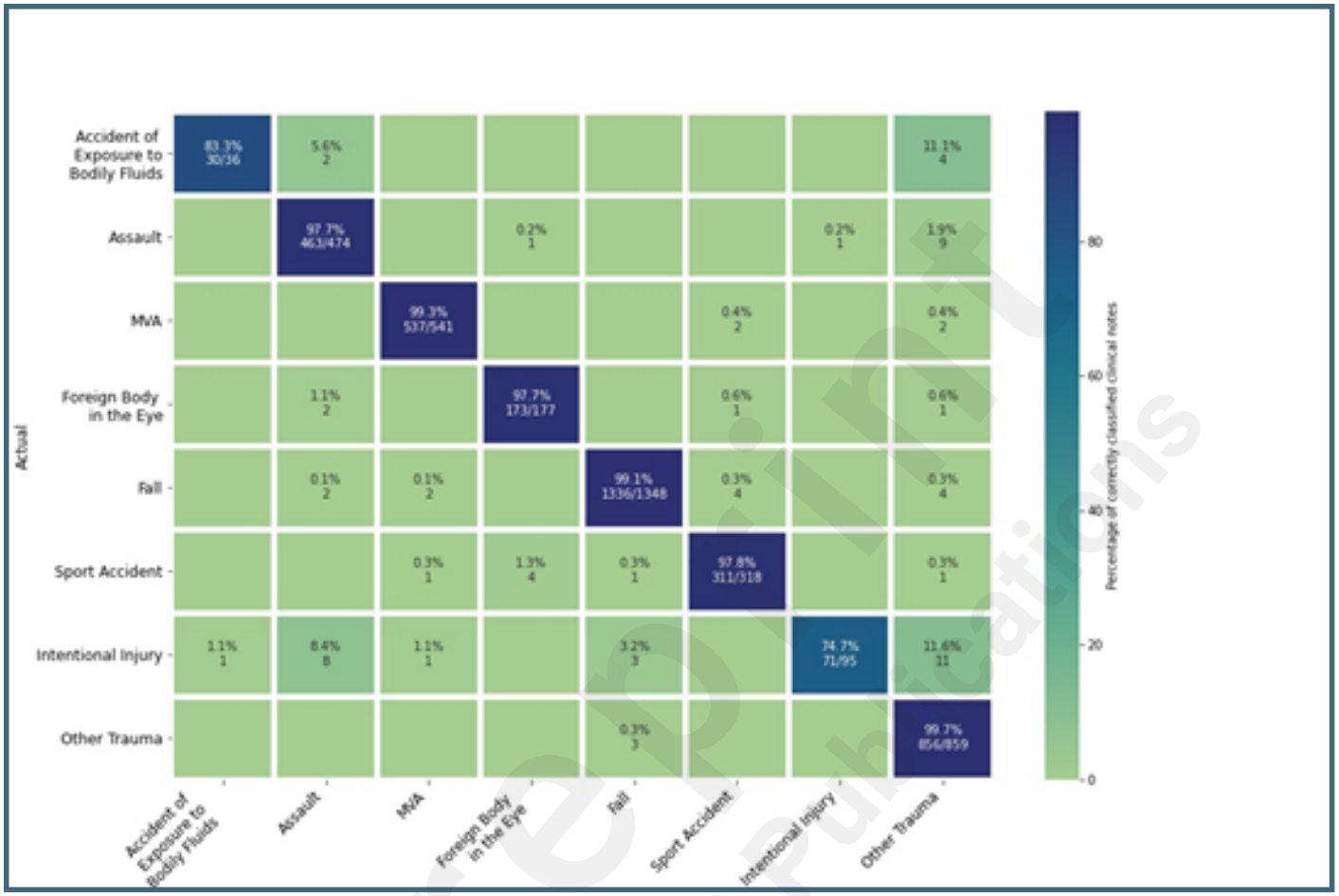
Patient emmené par les pompiers pour chute à domicile, TC sans PC. Sous kardegic. Chute mécanique, lésion du scalp.	Patient brought by the fire department for a fall at home with CT without LOC. Under kardegic. Mechanical fall scalp lesion.
AVP VL/VL, impact frontal, faible vitesse, airbag déclenché, pas de PC, pas PCI, pas dlr cou, pas céphalées	MVA LV/LV, frontal impact, low speed, airbag activated, no CT, no ILOC, no neck pain, no headache

Supplementary Figure 4. Average Macro-precision and Micro F1-score for each model for the test dataset without complex/ambiguous content in clinical notes.

Type of trauma	*N*	TF-IDF	CamemBERT	FlauBERT	BelGPT2	GPTanam
Accident of Exposure to Bodily Fluids	36	0.81	0.82	0.84	0.84	0.90
Assault	474	0.93	0.93	0.94	0.93	0.97
Motor Vehicle Accident	541	0.97	0.97	0.97	0.98	0.99
Foreign Body in Eye	177	0.80	0.85	0.83	0.83	0.97
Fall	1348	0.95	0.96	0.95	0.97	0.99
Sport Accident	318	0.89	0.91	0.91	0.93	0.98
Intentional Injury	95	0.79	0.80	0.75	0.81	0.85
Other trauma	859	0.84	0.86	0.86	0.89	0.98
Total	3848					
Micro average F1-score		0.904	0.921	0.918	0.932	0.981
Macro average precision		0.902	0.921	0.919	0.932	0.982

Supplementary Figure 5. Confusion Matrix for the GPTanam model on the complete test dataset. Ratio and percentage of correctly classified clinical notes per class are given. MVA: Motor Vehicle Accident.

Supplementary Figure 6. Confusion Matrix for the GPTanam model on the test dataset without complex/ambiguous content in clinical notes. Ratio and percentage of correctly classified clinical notes per class are given. MVA: Motor Vehicle Accident.



Supplementary Figure 1. Composite variable creation.