IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 22 November 2023, accepted 4 January 2024, date of publication 15 January 2024, date of current version 2 February 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3354705

RESEARCH ARTICLE

Pseudo-Labeling With Large Language Models for Multi-Label Emotion Classification of French Tweets

USMAN MALIK[1], SIMON BERNARD[1], ALEXANDRE PAUCHET[2], CLÉMENT CHATELAIN[2], ROMAIN PICOT-CLÉMENTE[3], AND JÉRÔME CORTINOVIS[4]

[1]Université de Rouen Normandie, LITIS UR 4108, F-76000 Rouen, France
[2]INSA Rouen Normandie, LITIS UR 4108, F-76000 Rouen, France
[3]Saagie, 76140 Rouen, France
[4]Atmo Normandie, 76000 Rouen, France

Corresponding author: Simon Bernard (simon.bernard@univ-rouen.fr)

This work is part of the CATCH research project, recipient of the RA-SIOMRI 2021 funding program, financed by the French National Research Agency (ANR) and the Normandy region under the ANR-21-SIOM-0011 grant.

ABSTRACT This study proposes a novel semi-supervised multi-label emotion classification approach for French tweets based on pseudo-labeling. Human subjectivity in emotional expression makes it difficult for a machine to learn. Therefore, it necessitates training supervised learning models on large datasets annotated by multiple annotators. However, creating such datasets can be costly and time-consuming. Moreover, aggregating annotations from multiple annotators to capture their collective emotional state adds complexity to the task. Semi-supervised learning techniques have shown effectiveness with limited datasets. Furthermore, Large language Models (LLMs), particularly Chat-GPT, have demonstrated superior annotation accuracy compared to annotations obtained from crowdsourcing platforms, when both are evaluated against expert-annotated data. This work introduces a novel approach for multi-label emotion classification of French tweets by leveraging pseudo-labels generated through Chat-GPT, a robust large language model. Using zero-shot, one-shot, and few-shot learning techniques, Chat-GPT annotates the unlabelled part of our dataset. These Chat-GPT-annotated pseudo-labels are then merged with manual annotations, facilitating the training of a multi-label emotion classification model via semi-supervised learning. Furthermore, within the context of our research, we present a new French tweet dataset, containing testimonials from people affected by an urban industrial incident. This dataset features 2,350 tweets, each manually annotated by three human annotators based on 8 pre-identified emotions. Benchmark results are presented for multi-label emotion classification models employing both fully supervised and semi-supervised approaches with pseudo-labeling. Our findings demonstrate the superiority of our approach for multi-label emotion classification over standard pseudo-labeling and an established baseline.

INDEX TERMS Multi-label emotion classification, semi-supervised learning, pseudo-labeling, Chat-GPT.

I. INTRODUCTION

Emotion classification in textual data involves identifying the latent emotional states, such as happiness, sadness, anger, fear, etc. in a text. This distinct analysis of emotions differs from sentiment analysis, which predominantly focuses on ascertaining the overall polarity of a text, encompassing positive, negative, or neutral sentiments [1]. In contrast, emotion classification aims to pinpoint and categorize the precise emotional expressions conveyed within a text, providing a more nuanced understanding of the underlying affective states [2].

Emotion classification is frequently approached as a multi-label classification problem involving two or more emotion labels [3], [4]. Although this strategy facilitates the detection of multiple emotional states in text, it also

The associate editor coordinating the review of this manuscript and approving it for publication was Seifedine Kadry.

introduces complexity to the task, due to the subjective nature of expressing emotions through language. In the context of emotion classification, individuals commonly employ identical words to convey distinct emotions or use different words to express a single emotion [5], [6]. To address this subjectivity, data-driven models for emotion classification necessitate datasets annotated by multiple annotators, thereby incurring significant costs and time requirements.

Furthermore, data annotated by multiple annotators rise the challenge of annotation aggregation. The inter-annotator agreement, which quantifies the level of consensus among annotators for a given dataset, is frequently observed to be low [7], [8]. Consequently, it is imperative to employ a suitable approach to aggregate annotations from multiple annotators, considering the specific requirements of the problem under consideration.

Semi-supervised approaches are often employed to address the challenge of limited datasets. Among these approaches, pseudo-labeling has demonstrated value to construct emotion classification models when confronted with limited datasets [9], [10]. Pseudo-labeling involves training a supervised learning model on an annotated dataset, using this model to predict labels for an unannotated dataset and subsequently incorporating these predictions, known as pseudo-labels, back into the training set to improve the model performance. This technique serves as a mean to increase the size of the training data and improve the overall classification accuracy. However, the success of pseudo-labeling depends on the reliability of the classifier trained on the annotated data, which in turn requires a sufficiently large and well-annotated training set. The inherent complexity of multi-label emotion classification further exacerbates the annotation costs.

An alternative method to augment the size of annotated data is to leverage pre-trained models, which are machine learning models trained on large datasets and are available for reuse. To this end, preliminary studies on the use of pre-trained large language models such as Chat-GPT for annotation, have shown promising potential [11]. This finding presents a prospect to leverage large language models to pseudo-label unannotated datasets, thereby enhancing overall classification performance across various tasks.

This article introduces a semi-supervised multi-label emotion classification approach for French tweets. The approach consists in generating pseudo-labels for unannotated data using Chat-GPT and merging the pseudo-labels with manual annotations to train a multi-label emotion classification model. In the applicative context of this study, a new publicly available dataset is presented, consisting of French tweets related to industrial accidents. The dataset is manually annotated by three annotators for eight distinct emotions and can be used for various emotion classification tasks. Detailed guidelines for annotation and an approach to aggregating annotations from multiple annotators are provided.

Benchmark study is presented for the multi-label emotion classification model trained on manually annotated tweets from the dataset. Additionally, results for models trained using pseudo-labeling techniques with data annotated from Chat-GPT using zero-shot, one-shot, and few-shot learning approaches are also included.

The two contributions of this article are the following.

- We propose a novel semi-supervised multi-label emotion classification approach that exploits pseudo-labelled data from Chat-GPT, annotated using zero-shot, one-shot, and few-shot learning.
- We present a publicly available,[1] manually annotated French tweets dataset for multi-label emotion classification, accompanied by detailed annotation guidelines and an annotation aggregation procedure. Benchmark results for multi-label emotion classification are also presented.

These contributions provide a road map to improve the multi-label emotion classification model with pseudo-labeling using LLM pre-trained model like chat-GPT. In this study, this is applied to AI-assisted crisis management in the context of an industrial accident in an urban environment. Further details about this applicative context are given in Section III.

This article is divided into 6 sections. Section II provides a literature review of the current state of the art in multi-label emotion classification. Section III presents the French tweet dataset for multi-label emotion classification. Section IV provides detailed explanation of the proposed approach. Experimental details and results are discussed in Section V. Finally, Section VI concludes this article and presents future directions of the study.

II. LITERATURE REVIEW

Multi-label emotion classification has been the target of research since the inception of natural language processing techniques. Over time, this area of study has experienced a significant upsurge in attention, driven notably by the introduction of transformer architectures and large language models. Nonetheless, the advancement of supervised learning approaches necessitates substantial annotated datasets, a resource-intensive and time-consuming endeavor. In response, recent efforts within the field of semi-supervised learning have been dedicated to addressing the challenges posed by limited datasets. This research effort resides at the intersection of two domains: multi-label emotion classification in text and semi-supervised learning, with a specific focus on the concept of pseudo-labeling.

In this context, this section provides an overview of existing approaches for multi-label emotion classification in texts. The challenges underlying this task, as well as the solutions available in the literature, are discussed, with a particular focus on semi-supervised learning and pseudo-labeling.

A. MULTI-LABEL EMOTION CLASSIFICATION

Historically, multi-label emotion classification has relied on heuristic approaches, including the use of emotion lexicons

[1] https://github.com/smnbrnrd/EmoDIFT

and manual feature extraction techniques [12]. These methods, whereas effective to some extent, often struggle to capture the complex nuances of emotions expressed in text. With the recent emergence of deep learning and the advent of attention mechanisms, the current state-of-the-art in multi-label emotion classification has shifted towards the adoption of deep learning techniques.

Multi-label emotion classification has leveraged convolutional (CNN) and recurrent (RNN) neural networks [12], [16]. While effective for short text sequences, their performance tends to suffer when handling longer sequences, often attributed to memory loss resulting from vanishing gradient problem [17], [18].

Graph Neural Networks (GNNs) have gained attention in recent research for multi-label emotion classification due to their ability to capture dependencies and interactions among emotional states, by leveraging structural relationships [19]. However, despite promising results, GNNs have limitations for this type of task, in terms of interpretability, scalability, generalization, and incorporation of external knowledge [19].

Attention mechanisms and transformer architectures have then emerged as powerful techniques to achieve state-of-the-art results in various natural language processing (NLP) tasks, including multi-label emotion classification [12], [20]. Transformers operate by employing self-attention mechanisms, capturing contextual relationships between tokens (words) in a sequence. This is accomplished through a mechanism that allows each word to attend to relevant parts of the input, resulting in a contextualized representation for each word. This attention-based approach has demonstrated exceptional performance, surpassing traditional CNN, RNN, and GNN models, especially in tasks involving long texts [21], [22], [23].

Unfortunately, despite promising results of deep learning for various text classification tasks, their performance often deteriorates when confronted with limited annotated data for training [24]. To alleviate this issue, transfer learning is commonly employed, where models pre-trained on large-scale datasets are fine-tuned on smaller annotated datasets [12]. However, transfer learning may not fully address the challenge of limited labelled data, as the models may still struggle to generalize well in the absence of sufficient labelled instances [25].

In the context of text classification, a huge amount of data is available on the internet, particularly on social media. However, this data is normally not annotated, and manual annotation can be expensive. One promising approach to tackle the limited annotated data challenge is the use of semi-supervised learning techniques [26]. Semi-supervised learning aims to leverage both labelled and unlabelled data during training, allowing the model to learn from the additional unlabelled instances. In the subsequent section, we will provide a brief overview of these semi-supervised learning techniques and their potential applications in addressing the challenges posed by limited annotated data in multi-label emotion classification tasks.

B. SEMI-SUPERVISED LEARNING FOR TEXT CLASSIFICATION

Semi-supervised learning is an approach that combines a small set of labelled data with a larger set of unlabelled data during model training [9]. In the context of text classification and more particularly multi-label emotion classification, semi-supervised learning techniques have shown promise in addressing the challenge of limited annotated data [27], [28]. Existing works divide semi-supervised learning approaches into four main categories: graph-based, unsupervised preprocessing, intrinsically semi-supervised approaches, and wrapper methods. [43].

Graph-based approaches for semi-supervised learning involve constructing a graph that represents the labelled and unlabelled data points, with edges denoting their similarity. Various similarity measures, such as cosine similarity or Euclidean distance, can be used for graph construction. By propagating the labels of the labelled data points along the graph edges, the labels of the unlabelled data points can be inferred [48]. This framework allows for leveraging the graph structure to enhance classification performance in a semi-supervised learning setting. Graph-based approaches have been widely used for semi-supervised text classification of news articles, social media texts, web pages, public sentiment, and scientific articles [49], [50], [51], [52], [53].

Unsupervised preprocessing techniques employ a two-step approach: unsupervised learning followed by supervised learning [43]. In the first step, an unsupervised algorithm extracts meaningful features from input data, capturing underlying patterns without relying on labelled information. Common techniques employed in this step include clustering, dimensionality reduction, and autoencoders. In the second step, preprocessed data is fed to a supervised algorithm, which maps preprocessed features to target labels using labelled data, optimizing model performance. Unsupervised preprocessing techniques such as *unsupervised feature extraction* [29], [30], [31] and *cluster-then label* [32], [33], [34] are widely employed for text classification tasks.

Intrinsically, semi-supervised approaches are designed to incorporate both labelled and unlabelled data during the learning process [35], [36], [37]. These methods leverage a combination of labelled and unlabelled loss terms to optimize the model performance. By jointly considering labelled and unlabelled data, the model can learn better representations and improve generalization. Techniques such as entropy minimization [38], consistency regularization [39], and generative models [40] are commonly used to effectively exploit unlabelled data. Intrinsically semi-supervised methods have proven effective in text classification scenarios with limited labelled data [41], [42].

Wrapper methods in semi-supervised learning involve a sequential process where a subset of labelled data is used for initial model training, followed by performance evaluation on a separate validation set. Subsequently, the trained model is employed to generate predictions on the unlabelled data, and

a subset of instances with high prediction confidence or low uncertainty scores is selected to augment the labelled set. This combined labelled data, comprising the original labelled data along with the newly selected instances, is then utilized to retrain the model [43].

Pseudo-labeling is one of the most commonly used wrapper methods for semi-supervised learning [44], [45]. Pseudo-labeling involves assigning temporary labels to the unlabelled data based on the predictions made by a pre-trained model. These pseudo-labels are then processed as additional labelled data and combined with the original labelled data for model training. The pre-trained model is often trained using only the labelled data and is used to make predictions on the unlabelled instances. Pseudo-labeling has emerged as a prominent technique for enhancing the performance of text classification tasks, particularly in scenarios with limited datasets [10], [54], [55], [56].

The existing semi-supervised approaches have their own advantages and drawbacks. While graph-based approaches for semi-supervised learning do offer advantages by utilizing the relationships between data points, they do come with certain limitations. These methods can become computationally intensive, which can become an obstacle when working with large datasets. Moreover, they demand careful consideration in terms of selecting appropriate similarity measures and constructing graphs [57]. Alternatively, unsupervised preprocessing methods heavily rely on the performance of the unsupervised learning algorithm for feature extraction. This can introduce challenges in interpreting the outcomes effectively. Furthermore, intrinsically semi-supervised methods that inherently integrate both labelled and unlabelled data necessitate the tuning of numerous parameters, such as loss terms and regularization techniques. Often, such parameter tuning requires a substantial annotated dataset to yield meaningful results.

Among the approaches discussed in this section, pseudo-labeling has emerged as a highly effective and popular technique in semi-supervised learning for several reasons. Firstly, pseudo-labeling is known for its computational efficiency, as it does not require complex graph construction or feature extraction algorithms [46]. Secondly, pseudo-labeling offers flexibility by being compatible with various models and datasets. It can be seamlessly integrated with different machine learning algorithms and adapted to different domains and problem settings [45]. Furthermore, pseudo-labeling has proven its effectiveness by achieving state-of-the-art performance on various semi-supervised learning benchmarks [47]. Overall, pseudo-labeling stands out as a preferred choice in semi-supervised learning due to its simplicity, efficiency, flexibility, and remarkable performance in various domains.

In this work, we propose a semi-supervised pseudo-labeling approach for multi-label emotion classification. In this context, the next section reviews recent pseudo-labeling approaches for text classification.

C. PSEUDO-LABELING FOR TEXT CLASSIFICATION

Pseudo-labeling involves predicting labels for unannotated data using a model trained on annotated data. These predicted labels, known as pseudo-labels, are then merged with the annotated training set to retrain the model. Pseudo-labeling in most cases, improves the overall classification performance of the model.

Existing works on semi-supervised learning with pseudo-labeling can be grouped into three categories: traditional pseudo-labeling, lexicon-based pseudo-labeling, and pseudo-labeling with data augmentation.

Traditional pseudo-labeling approaches involve predicting labels for unannotated data using a model trained on annotated data. These predicted labels, known as pseudo-labels, are then merged with the annotated training set to retrain the model. To enhance the effectiveness of this process, traditional pseudo-labeling techniques commonly incorporate mechanisms designed to identify the most reliable labels. By doing so, they effectively filter out noisy pseudo-labels that have the potential to undermine the performance of models trained using such pseudo-labelled data. To this end, Li et al. propose S2TC-BDD (Semi-Supervised Text Classification with Balanced Deep Representation Distributions) [54]. S2TC-BDD achieves improved accuracy in predicting pseudo-labels by estimating variances over both labelled and pseudo-labelled texts. Mekala et al. propose LOPS (Learning Order Inspired Pseudo-Label) for weakly supervised text classification [59]. They leverage the learning order of samples to estimate the probability of incorrect annotation, based on the hypothesis that models prioritize memorizing samples with clean labels before those with noisy labels.

Lexicon-based pseudo-labeling involves a lexicon or a predefined set of rules to allocate labels to unlabelled data. This strategy exploits the lexicon to recognize keywords or patterns within the text that correspond to particular labels. Subsequently, these identified patterns and keywords guide the assignment of labels to the unannotated data. Hwang & Lee introduce a lexicon-based pseudo-labeling method using explainable AI (XAI) for sentiment analysis [55]. The method generates a lexicon of sentiment words based on explainability scores and calculates the confidence of unlabelled data using this lexicon. Jimenez et al. propose an approach that leverages named entity recognition (NER) based lexicon and subjectivity measures to discern news from non-news content on websites [61].

Data Augmentation-Based Pseudo-Labeling encompasses generating synthetic labelled data using data augmentation techniques and then using a classifier to filter the instances with high-confidence pseudo-labels. The filtered instances are merged with the original training data for pseudo-labeling. As an example, Yang et al. present STCPC (Small-sample Text Classification model using Pseudo-label fusion Clustering) that combines pseudo-labeling and data augmentation techniques to improve text classification with limited

labelled data [10]. Wang & Zou combine pseudo-labeling and the MixMatchNL data enhancement technique [58] to predict label data and address the imbalance in short text datasets [56]. Anaby-Tavor et al. propose LAMBADA (Language-Model-Based Data Augmentation) a GAN-based approach for data augmentation in text classification [62]. They refine a pre-trained GAN to create synthetic labelled data and merge it with the original dataset for pseudo-labeling.

In addition, it is observed that while some studies have focused on sentiment detection in public reviews (e.g., Amazon, Yelp, and IMDB), there is a scarcity of research addressing multi-label emotion classification. The limited exploration of pseudo-labeling for text multi-label emotion classification can be attributed to the challenge of subjective and overlapping labels. Successfully applying pseudo-labeling to these intricate tasks necessitates larger training sets annotated by multiple annotators, which may be challenging to acquire due to time and cost constraints, thus limiting the effectiveness of existing pseudo-labeling techniques.

This research paper aims to address the limitations of current semi-supervised learning approaches when applied to complex problems such as multi-label emotion classification. Specifically, we propose a novel semi-supervised learning technique based on traditional pseudo-labeling that leverages Large Language Models (LLMs), such as ChatGPT, to predict labels for unlabelled data. These pseudo-labels are then integrated into the training process to enhance learning performance. We refrain from utilizing the lexicon-based pseudo-labeling approach due to its reliance on constructing a lexicon linked to a specific domain, which may lead to generalization issues. Our intention is to introduce a more adaptable approach that can be employed for text classification tasks beyond the scope of our current study. Furthermore, pseudo-labelling via data augmentation is not appropriate for our particular problem, as generating ''fake yet realistic'' tweets that accurately capture different human emotions is an open problem that may be an interesting direction for future research, but is outside the scope of this work.

Among traditional pseudo-labeling approaches, we have chosen the LOPS model by Mekala et al. [59] as the state-of-the-art baseline among traditional pseudo-labeling approaches, owing to its generalization capabilities and less dependency on hyperparameters. We opt not to consider the method by Li et al. for benchmarking due to its substantial reliance on computationally expensive hyperparameters.

The next section provides a comprehensive overview of the dataset employed in our research work for multi-label emotion classification in French tweets. We describe the process of data collection, the annotation protocol employed, the inter-annotator agreement, and the post-processing technique implemented on the collected data.

III. MULTILABEL EMOTION CLASSIFICATION DATASET OF FRENCH TWEETS

This research is part of a larger project that intends to develop a model capable of detecting public emotions from French tweets in real-time during industrial accidents. The ultimate goal is to address both the immediate and long-term impacts of the disaster and adapt crisis management strategies based on public emotions. However, developing such a model requires a dataset containing French tweets from the public during and after industrial accidents. To the best of our knowledge, no existing French dataset contains tweets annotated with multiple emotion labels. This motivates us to develop a first-of-its-kind multi-label emotion-annotated dataset of French tweets.

The Lubrizol factory fire, which occurred on September 26, 2019, in Rouen, France, serves as a notable example of an industrial accident. The incident resulted in significant damages due to the combustion of chemicals and fuel additives, leading to a substantial response on social media platforms [14], [65], [66], [67]. In this study, we collected tweets related to this factory fire incident, manually annotated them with emotions, and post-processed them to train a semi-supervised multi-label emotion classification model. This dataset stands as a pioneering contribution within the realm of multi-label emotion classification datasets for French tweets. The next section describes the data collection and annotation protocol.

A. DATA COLLECTION AND ANNOTATION

The dataset consists of 90,496 tweets from 21 September 2019 to 30 December 2020, all of which mention the term ''Lubrizol''. These tweets originate from a diverse range of sources, encompassing both individuals and organizations. For this work, we consider that emotions are expressed in tweets from the public only. So our first task is to filter the tweets posted by organizations, public authorities, the press, and so on. This filtering task is not trivial, but presents far fewer difficulties than our emotion detection task, since it involves binary classification for which the two classes are unambiguous: population or not. This is why we simply manually annotated 300 tweets in order to fine-tune and test a pre-trained CamemBERT model [75], which achieves a F1 score of 0.90 on the test set. From the remaining unlabelled dataset, predictions with a probability threshold > 0.80 are retained and the remaining instances are filtered out. The main goal of this filtering is to exclude as many press-related tweets as possible from our dataset, in order to reduce the potentially significant imbalance between neutral tweets and those that convey emotions. This is the reason why we do not seek to achieve the highest possible accuracy for this filtering step. At the end of this filtering stage, a set of 12,508 tweets from individuals is obtained. From this subset, 2,350 tweets are randomly selected and manually annotated by a total of 11 human annotators, including university professors, researchers, engineers and students, all of whom are native French speakers.

The annotators are instructed to read and evaluate tweets, with the goal of identifying a maximum of three emotions from a predefined set of emotion labels. If the emotions

expressed within the tweet belong to multiple emotion registers, the annotator is required to rank them in order of importance. This order is based on the predominance of expressed emotions within the tweet. Specifically, the label associated with the dominant emotion is assigned a value of 1, while the second and third most dominant emotions are assigned values of 2 and 3, respectively. The ultimate goal of the annotation is to identify expressions of opinion, whether they are explicit (*I'm scared*), or implicit (*Those are hydrocarbons in there, that's dangerous*), taking only the semantic information into consideration.

The ranking approach is utilized to annotate tweets with emotions, thereby offering a solution to a number of problems such as 1) multi-class classification where the emotion possessing the highest rank is treated as the target class; 2) multi-label classification where emotion ranks can be treated as the target labels, or alternatively, ranks can be converted into one-hot encoded target labels; and 3) multi-label regression problem where ranks correspond to the degree of dominance of emotions.

The negative emotion labels are selected from the second tier of Plutchik's model of emotions, due to their balanced representation between the top-tier and low-tier emotions [68]. This choice allows to capture the nuanced nature of emotions within the broader emotional framework proposed by Plutchik. Given the nature of our case study, 6 emotions are selected from this categorization model, namely *anger, disgust, fear, surprise, sadness,* and *mistrust*.[2] In addition, *irony* has been included as an optional label since irony has been shown to be an important clue for emotion classification [69]. Furthermore, the labels *neutral* and *inexploitable* have been added to allow annotators to identify tweets that do not express aany particular emotion or that are not relevant to our case study (for example, a tweet from a news media that has not been filtered in the preprocessing phase).

As discussed, emotion annotation poses a complex challenge owing to the subjective nature of emotional expression hindering consensus among annotators. Moreover, the brevity of Tweets and the common use of colloquial language contribute to the ambiguity of the emotions expressed. To examine these assumptions and determine the level of agreement among annotators in our dataset, the following section presents an analysis of inter-annotator agreement (IAA) for our dataset.

B. INTER ANNOTATOR AGREEMENT

The inter-annotator agreement (IAA) refers to the overall degree of agreement between multiple annotators for the underlying annotations. Among statistical approaches for IAA, we select Krippendorff's Alpha [71], and Fleiss' Kappa [72]. The values for these measures range from 0 to 1, where 0 signifies no agreement and 1 represents complete agreement. We select these metrics since they satisfy the two annotation criteria of our annotation protocol: (1) they can be used to find IAA between more than two annotators, and (2) they can find IAA for ranked data.

TABLE 1. Inter-annotator agreement results (Kripp = Krippendorffs, A = Alpha, K = Kappa).

Emotion	Kripp's A	Fleiss' K
Anger	0.2247	0.2245
Disgust	0.1410	0.1409
Fear	0.2574	0.2572
Mistrust	0.1946	0.1945
Sadness	0.1425	0.1424
Surprise	0.1093	0.1092
Irony	0.2662	0.2661
Neutral	0.1631	0.1630
Inexploitable	0.0982	0.0981
Mean	0.1774	0.1773

The IAA results for Krippendorff's Alpha and Fleiss' Kappa are presented in Table 1. Krippendorff's Alpha exhibited a mean value of 0.1774, suggesting a notably low level of agreement among the annotators [73]. Similarly, Fleiss' kappa values aligned closely with Krippendorff's alpha, averaging at 0.1773 across all emotions, indicating minimal to no agreement [74].

A close analysis of IAA reveals that specific emotions, namely *irony, fear*, and *anger*, exhibit Krippendorf's Alpha and Fleiss' Kappa values ranging from 0.20 to 0.266, which suggest a moderate level of agreement, indicating a fair degree of consensus among multiple annotators for these particular emotions.

In addition to statistical measures, we also conduct heuristic tests to further investigate the IAA in our study. The results indicate that only 4.83% of tweets exhibit complete agreement among the three annotators across all ranks and all emotions. For 5.48% of the tweets, the annotators agree on the emotion labels but differ in their assigned ranks. It is noteworthy that these percentages are relatively high when considering the agreement between any two annotators out of the three. Breaking down the results by individual ranks, the annotators exhibit the highest agreement of 16.74% for rank 1, followed by 1.44% for rank 2. However, for rank 3, there is no agreement observed among all three annotators. Further investigation of the agreement for rank 1 reveals that *anger, mistrust*, and *irony* demonstrate the highest agreement percentages of 40%, 23%, and 12%, respectively. This sheds light on the emotions that tend to exhibit stronger consensus among annotators for rank 1.

The statistical and heuristic analysis of IAA exposes the subjective aspect of human emotion expression. This finding underscores the significance of employing a reliable aggregation technique that can effectively capture the emotional state reflected by multiple annotators. The following section details the post-processing approach utilized in this study for that purpose.

[2]The original dataset is in French language with the following emotions labels: *colère* (anger), *dègout* (disgust), *peur* (fear), *méfiance* (mistrust), *tristesse* (sadness), *surprise* (surprise), *ironie* (irony), *neutre* (neutral).

C. POST-PROCESSING ANNOTATIONS

In this study, we focus on the multi-label classification of emotions in tweets. Annotating tweets with emotion ranks enables us to estimate a degree of presence for each predefined emotion. This also allows to reduce annotator bias, which is relatively high given the subjectivity of the task. The next step is therefore to merge the three human annotations of each tweet to obtain a one-hot encoded vector, which identifies only the emotions that have reached a certain degree of prominence. It is worth noting that such annotations would enable us to approach the problem from other learning paradigms: multi-class classification, multi-output regression or even learning to rank. The funded project in which this study is part focuses on the multi-label classification of emotions in tweets.

Voting and mean calculation are common approaches to aggregate annotations from multiple annotators [13]. However, they are unsuitable for ranked annotations due to (i) higher mean values for lower ranks, and (ii) the lack of ranked information in the default voting mechanism. For example, two annotators assigning rank 1 to a text is not equivalent to one assigning rank 1 and the other assigning rank 2 or 3.

This problem is addressed in [70] by introducing an approach that transforms ranked annotations from multiple annotators into normalized degrees of class membership. These values range from 0 to 1 and sum up to 1, where higher values correspond to a more pronounced presence of emotion. They employ a voting mechanism to select tweets in which a minimum of two annotators allocate any rank to at least one label. For the selected tweets, the normalized degrees of class membership are calculated to derive the final aggregated values. The approach was evaluated on the TREMoLO dataset [70], which contains tweets annotated with language registers indicating whether a tweet is formal, neutral, or casual. These label annotations are less subjective compared to emotion classification, resulting in a less complex problem setting. For this reason, in our study, we have adopted and modified this method for our problem, as explained below.

First, ranks are transformed into degrees of belonging using the formula in [70]. For the sake of brevity, we do not give the formula here, but refer the reader to [70] for the details of this calculation. However, tweets are not subsequently filtered using the voting mechanism, since it results in the removal of a large number of tweets in our dataset due to the low IAA. Moreover, tweets with an *unexploitable* label exceeding 0.5 are eliminated from the dataset based on the consensus of annotators regarding their informational deficiency. Furthermore, due to the larger number of annotators and output labels, as well as low IAA, the individual degrees of belonging for specific labels in our dataset become significantly small. As a result, the thresholding approach (>0.5) employed in [70] for converting degrees of belonging to one-hot encoded vectors is inadequate for emotions in our case. To address this, we adopt an alternative approach by selecting the top N emotion labels with the highest degrees of belonging. Our choice of $N = 3$ is based on two reasons: (i) annotators were instructed to rank up to three emotion labels, and (ii) after conducting tests with different N values, we determined that $N = 3$ yields the most optimal results. It is important to note that in the case of equal values for the degrees of belonging, both emotion labels are selected, sometimes resulting in the selection of 4 labels.

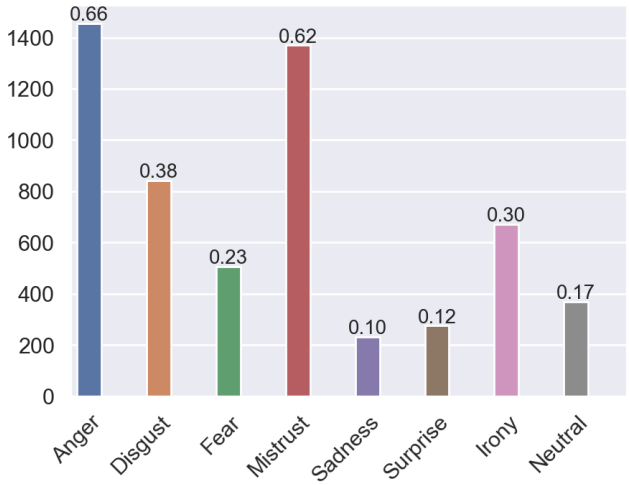


FIGURE 1. Emotion distribution by tweet percentage. The y-axis represents the number of tweets. Bar values represent percentages.

From the 2,350 manually annotated tweets, 2,215 are retained after excluding the tweets labelled as *unexploitable*. Figure 1 illustrates the distribution of emotions within the annotated tweets, providing a visual representation of their prominence. The analysis reveals that the two labels *anger* and *mistrust* are particularly present in the datasets, namely for 66% and 62% of the 2,215 tweets, respectively. On the other hand, *sadness* and *surprise* exhibit considerably lower representation, accounting for only 10% and 12% of the tweets, respectively. This highlights an additional, but expected, difficulty inherent to our case study, namely that all the selected emotions are expressed in very unequal ways in our datasets, resulting in highly imbalanced classes.

The low IAA values and the highly imbalanced nature of the dataset emphasize the necessity for large annotated datasets to develop robust classification models. In scenarios like this, pseudo-labeling approaches that rely on a model trained with a limited number of manual annotations for pseudo-label prediction exhibit poor performance [80]. The primary cause of the unsatisfactory performance stems from the fact that datasets with minimal IAA for multi-label emotion classification underscore the inherent subjectivity of the task. When working with limited data, capturing this subjectivity becomes challenging. As a consequence, pseudo-labels generated through training on such data tend to be erroneous, contributing additional noise to the training process. This, in turn, leads to poor performance, even when compared to a fully supervised model.

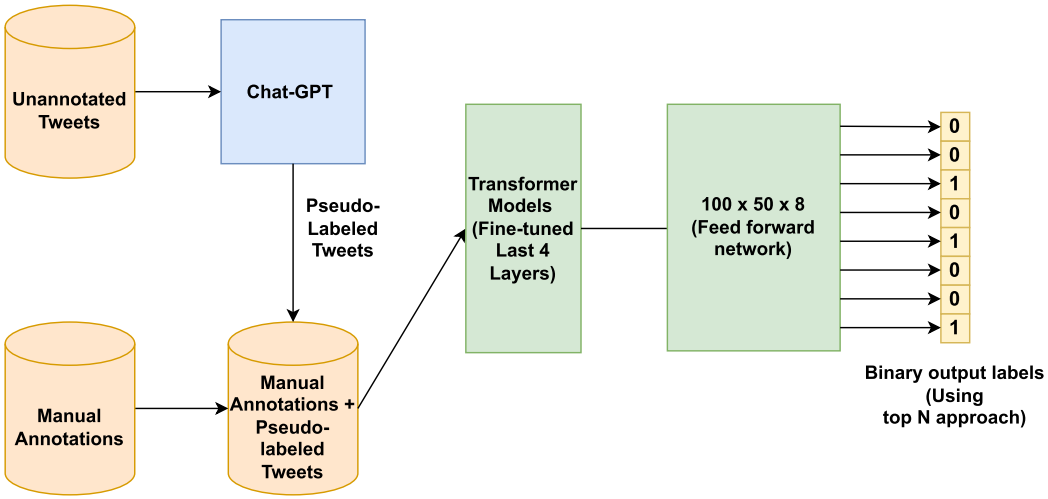


FIGURE 2. Semi-supervised training using pseudo labels from Chat-GPT.

A potential solution is to employ pre-trained large language models that have been trained on extensive datasets sourced from various origins. These models are better positioned to capture the inherent subjectivity of human nature. This can potentially address the challenge of human subjectivity and lack of data by generating more informed pseudo-labels that when combined with the manual training set result in improved model performance.

In this study, we propose a semi-supervised multi-label emotion classification approach that uses pseudo-labels from a pre-trained large language model (Chat-GPT). The following section outlines the proposed approach.

IV. PROPOSED APPROACH

Existing works have demonstrated Chat-GPT ability for data annotation [11]. In the approach we propose in this study, we first annotate a set of unannotated tweets from a dataset using Chat-GPT and then use the Chat-GPT pseudo-labels combined with manual annotations in the training set to train multi-label emotion classification models.

Chat-GPT prompts offer a high degree of flexibility, accommodating a wide range of textual inputs. When it comes to annotation, the outcome is significantly influenced by the specific prompt employed. Interestingly, even two prompts sharing the same semantic meaning may yield distinct annotations. This emphasizes the crucial role of careful prompt construction. Although dealing with the intricacies of prompt engineering is beyond the scope of this study, three different prompting approaches are explored for Chat-GPT-based pseudo-labeling, with three different levels of instructions:

- **Zero-shot**: The prompt instructs Chat-GPT to select emotion labels from a predefined label set without providing any examples from the manually annotated dataset.
- **One-shot**: The prompt instructs Chat-GPT to select emotion labels from the label set, using one example for each emotion from the manually annotated dataset.
- **Few-shots**: The prompt instructs Chat-GPT to select emotion labels from the label set, using two examples for each emotion from the manually annotated dataset.

For the sake of reproducibility, we provide in Appendix VI the exact prompts we used in our experiments for all three cases.

Using the above three approaches, a total of 2,800 additional tweets are annotated using Chat-GPT. To ensure high-quality annotations, tweets where Chat-GPT predicted *inexploitable* are removed from the dataset. As a result, the zero-shot, one-shot, and few-shot approaches retained a total of 2,767, 2,757, and 2,775 tweets respectively. Consequently, for the sake of comparison, the three pseudo-labelling approaches have been applied on the same unlabelled tweets, but have been separately merged with the manual annotations, resulting in three different training sets.

Figure 2 contains an overview of the proposed approach. The dataset is divided into unannotated tweets and manually annotated tweets. In the first step, unannotated tweets are pseudo-labelled with Chat-GPT using one of the three aforementioned approaches. The pseudo-labelled tweets are then merged with the manual annotations. The merged datasets containing pseudo-labelled and manually annotated tweets are then passed to transformer models for fine-tuning.

In this study, two variants (base and large) of the two most commonly used French language transformer models are tested for text embedding: CamemBERT [75] and FlauBERT [76]. These models are preferred since they have been specifically trained on French and have demonstrated state-of-the-art results for classification of French texts [78]. In order to also compare with a multilingual transformer model, experiments are complemented with RoBERTa [77]. It is known that fine-tuning this type of Large Language

Models is complex in a low-data regime [81]. It is known that fine-tuning this type of Large Language Models is complex in a low-data regime [81], [82]. In our experiments, we used the popular solution of freezing the first layers of the network and fine-tuning only the last four layers, as this proved empirically to perform best in our context.

The output from the transformer models is passed to two dense neural network (DNN) layers of sizes 100 and 50 (with ReLU activation functions) respectively, followed by an output layer of size 8, corresponding to 8 emotion labels in the dataset. The sigmoid activation function is used in the output layer returning values between 0 and 1 which corresponds to the probability of emotions present in the input tweet. We opted not to employ the softmax activation function in the output layer since objective does not involve identifying the distribution of emotions within a tweet. Instead, our focus is to detect in a tweet the presence of each emotion independently.

Since the target emotion labels are binary values, we employ a binary cross-entropy loss function for each of the eight emotion categories. Formally, the binary cross-entropy loss function for the *i-th* emotion is expressed as:

$$L_i(y_i, \hat{y}_i) = -(y_i * log(\hat{y}_i) + (1 - y_i) * log(1 - \hat{y}_i)) \quad (1)$$

where y_i is the binary target label of the *i-th* emotion and $\hat{y}_i$ is the corresponding prediction of our model. To assess the overall performance of the model, the individual binary losses for all $m = 8$ outputs are aggregated through a simple arithmetic mean, yielding the overarching loss metric:

$$\mathcal{L} = \frac{1}{m} \sum_{i=1}^{m} L_i(y_i, \hat{y}_i) \quad (2)$$

Finally, to ensure the consistency between the post-processed tweet labels and the output labels, the top 3 probability values given by the model are set to 1, while the remaining of the values are set to 0. Since our dataset is imbalanced, both in terms of emotions and ranks, we utilize the F1 measure as evaluation criterion, as it is suitable for imbalanced data classification models [79].

The proposed approach boasts simplicity in its implementation and demonstrates strong potential for generalization across diverse problem scenarios and case studies. Adapting this approach to a different problem setting is a straightforward process, necessitating only the substitution of a distinct dataset and a minor adjustment in the Chat-GPT prompts.

The following section explains the experiments and results of the proposed approach.

V. EXPERIMENTS AND RESULTS

The four sets of experiments performed in this study are the following ones:

- Fully supervised learning (baseline)
- Pseudo-labeling using Chat-GPT (proposed approach)
- Standard pseudo-labeling
- The state-of-the-art LOPs method from [59]

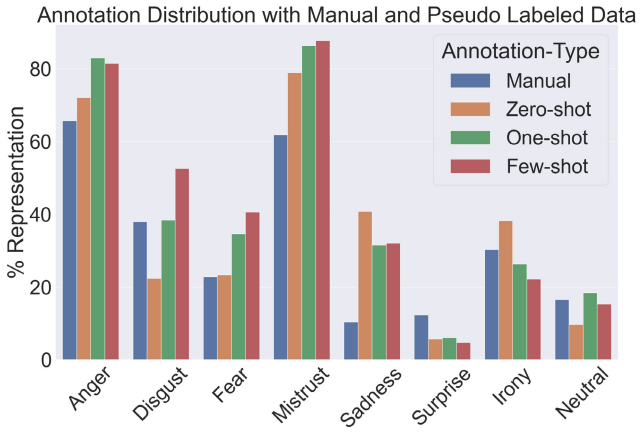


FIGURE 3. Annotation distribution with manual and pseudo labelled data.

The above experiments are performed on the exact same dataset where the same set of unlabelled tweets are pseudo-labelled using the proposed approach, the standard pseudo-labeling approach, and the LOPS model.

The experimentation process for the proposed approach follows the traditional machine-learning pipeline. The K-fold cross-validation approach (K = 5) is used to randomly split the 2,215 post-processed manual annotations into train-test sets. The pseudo-labels from Chat-GPT are merged with the manually annotated training set. The merged annotations are passed to transformer models (discussed in the previous section) with additional dense layers. Details of the model architecture and training hyperparameters are given in Table 2. These values were first set based on the recommendations given in [60], then empirically adjusted to obtain better results on our problem.

For a fully supervised learning baseline, a process similar to our proposed approach is adopted, however, in this case, only manual annotations are used for training the proposed model. For the standard pseudo-labeling baseline, a fully supervised model is exploited to make predictions on the unannotated tweets, which are then merged with the manual annotations to make predictions on the test set. The model architecture and hyper-parameters for the fully supervised and standard pseud-labeling approach remain the same as mentioned in Table 2.

Figure 3 depicts the emotion distribution for the Chat-GPT annotated tweets 3. The figure shows that except concerning *surprise*, the distribution of all the remaining emotions has increased in the Chat-GPT annotations. This shift in distribution can potentially be attributed to several factors. One possible explanation is the inherent nature of Chat-GPT training data, which includes a wide range of conversational content from various sources. It is highly likely that Chat-GPT encounters training data where *surprise* is less prevalent compared to other emotion labels. As a result, the model annotates fewer tweets with *surprise*. Moreover, the discrepancy could be due to differences in

TABLE 2. Model architecture and training parameters.

Model and Parameter Types	Description
Pretrained Models	TensorFlow versions of the pre-trained transformed models from Hugging Face are used for training. Only the last 4 layers of the models are fine-tuned. The last hidden state of the output is used as the input to the downstream deeply connected layers.
Token Sizes	512
Additional Layers	Two deeply connected layers of sizes 100 and 50, followed by an output layer of size 8 appended to the output of Hugging Face Models.
Drop out	Dropouts of 0.25 are used for the pre-trained model and the first deeply connected layer. A dropout of 0.15 is used for the second densely connected layer.
Activation Functions	Relu activation functions are used for the first two deeply connected layers. A sigmoid function is used for the final output layer.
Optimizer & Learning Rate	Adam optimizer with a learning rate of $5e^{-5}$
Loss Function	Binary Cross Entropy Loss
Training Epochs & Batch Size	A total of 30 epochs and a batch size of 8 with early stopping after 5 epochs if F1 score does not improve further. The best model is chosen for prediction on the test set.

TABLE 3. Mean F1 scores and standard deviations for multi-label emotion classification with Chat-GPT annotations and with baselines and LOPS approach. CGPT = Chat-GPT, PL = Pseudo-labeling.

Model	CGPT Zero-shot	CGPT One-shot	CGPT Few-shots	Standard PL	LOPS	Fully Supervised
CamemBERT-Large	0.6370 ± 0.0021	0.6649 ± 0.0051	$\mathbf{0.6662 \pm 0.0042}$	0.6310 ± 0.0121	0.6523 ± 0.0073	0.6538 ± 0.0255
CamemBERT-Base	0.6121 ± 0.0081	0.6289 ± 0.0281	0.6274 ± 0.0153	0.6146 ± 0.0413	0.6310 ± 0.0179	0.6201 ± 0.0092
FlauBERT-Large	0.5920 ± 0.0241	0.6019 ± 0.0075	0.6042 ± 0.0195	0.5812 ± 0.0310	0.5890 ± 0.0012	0.5901 ± 0.0077
FlauBERT-Base	0.5812 ± 0.0043	0.5875 ± 0.0724	0.5901 ± 0.0043	0.5798 ± 0.0015	0.5890 ± 0.0241	0.5870 ± 0.0085
RoBERTa-Large	0.5755 ± 0.0167	0.5770 ± 0.0024	0.5846 ± 0.193	0.557 ± 0.0042	0.5810 ± 0.0031	0.5775 ± 0.0087
RoBERTa-Base	0.5810 ± 0.0043	0.5809 ± 0.0081	0.5943 ± 0.0062	0.5802 ± 0.0073	0.5874 ± 0.0141	0.5821 ± 0.0110

the annotation process. Human annotators might interpret and label emotions differently from how the model does, leading to variations in distribution. In addition, Chat-GPT responses might be influenced by prompts provided during annotation.

Table 3 presents the overall results of the models trained using the proposed pseudo-labeling approach with Chat-GPT using zero-shot, one-shot, and few-shot annotations. The table also depicts results obtained via (i) standard pseudo-labeling, (ii) the LOPs method, and (iii) fully supervised training. The results indicate that the CamemBERT-Large model with Chat-GPT few-shot pseudo-labeling achieves the highest F1 score (0.6662) surpassing the performance obtained via standard pseudo-labeling (0.6310), the LOPS method (0.6523), and the fully supervised approach (0.6538). It is worth noting that the CamemBERT-Large model always yields the best performance for all the methods used in the experiment. Similarly, the Chat-GPT few-shot pseudo-labeling method is also always the most accurate, whatever the language model used.

The possible reason for the CamemBERT-Large model better performance compared to the FlauBERT model could be that FlauBERT is larger than CamemBERT and the complexity of FlauBERT may require a more extensive dataset for optimal fine-tuning. However, the comparison between CamemBERT-large and CamemBERT-small shows the opposite behavior where CamemBERT-large, despite being a larger model, results in better performance than CamemBERT-small. The reason for this behavior could be that CamemBERT-large is a compromise between FlauBERT and CamemBERT-small models. While it is less complex than FlauBERT, it is complex enough to grasp the intricacies of our dataset. Furthermore, CamemBERT better performance compared to RoBERTa could be attributed to RoBERTa multilingual nature encompassing a broader range of languages, which could lead to a dilution of its capacity to grasp the intricacies of any particular language such as French. Nevertheless, performance comparisons between various large language for specific tasks is open to further research.

TABLE 4. Mean F1 scores and standard deviations for individual emotions using semi-supervised pseudo-labeling with Chat-GPT few-shots annotations and LOPS model.

Emotion	CGPT Few Shots	Fully Supervised
Anger	0.7932 ± 0.0171	0.7901 ± 0.0100
Disgust	0.5872 ± 0.0105	0.5851 ± 0.0072
Fear	0.5732 ± 0.0087	0.5121 ± 0.0134
Mistrust	0.7710 ± 0.0174	0.7421 ± 0.0092
Sadness	0.4652 ± 0.0072	0.4142 ± 0.0172
Surprise	0.4731 ± 0.0189	0.5105 ± 0.0127
Irony	0.6432 ± 0.0075	0.6137 ± 0.0111
Neutral	0.4534 ± 0.0271	0.3752 ± 0.0192

In addition, it is worth noting that the difference in results between the one-shot pseudo-labeling and the few-shot pseudo-labeling approaches is very small. However, when compared to manual annotations, the results obtained through zero-shot pseudo-labeling exhibit inferior performance. This result depicts that while a single training example might be enough to achieve the best results, a Chat-GPT prompt without any example from the training set leads to inferior performance on the test.

TABLE 5. Prompts used for pseudo-labeling with Chat-GPT using zero-shot, one-shot, and few-shot annotations. Tweets are annotated via OpenAI's *GPT-3.5 Turbo* model with default settings except for the temperature attribute which is set to 0 to avoid randomness. OpenAI's Python API is used to make calls to the model. French emotion labels are used in the prompt.

Annotation Type	Example Prompt	Description
Zero-shot	You are a French linguist expert who annotates French tweets with emotions. Pick 4 or less the most likely emotions from the emotions list to annotate the input French tweet. **input french tweet** = "C'est l'équivalent d'une marée noire, mais sur terre. Le toit en amiante a été pulvérisé, les gens ont respiré les particules… C'est une vraie catastrophe écologique et sanitaire....." **emotions list** = Colere, Degout, Peur, Mefiance, Tristesse, Surprise, Ironie, Neutre, Inexploitable. The response text should only contain a comma-separated list of up to 4 most likely emotions present in the input French tweet. The emotions should be selected from the emotions list. If you can't find an emotion from the emotions list, simply return the value "Inexploitable"	The prompt entails submitting a tweet to Chat-GPT for annotation, with specific instructions to select a maximum of 4 emotions from a predefined list of 8 emotions used in manual annotations. Importantly, the prompt does not include any samples from the training set. Additionally, in scenarios where none of the emotions are discernible within the tweet, the expected outcome from Chat-GPT is *inexploitable* (unexploitable). Tweets categorized under the label *inexploitable* are excluded from the subsequent process of semi-supervised pseudo-labeling, resulting in a total of 2767 tweets for zero-shot pseudo-labeling.
One-Shot	You are a French linguist expert who annotates French tweets with emotions. Pick 4 or less most likely emotions from the emotions list to annotate the input French tweet. As examples use tweet examples and corresponding emotions from the training samples list: **input french tweet** = "C'est l'équivalent d'une marée noire, mais sur terre. Le toit en amiante a été pulvérisé, les gens ont respiré les particules… C'est une vraie catastrophe écologique et sanitaire....." https://t.co/HNNfvQmS3P. **emotions list** = Colere, Degout, Peur, Mefiance, Tristesse, Surprise, Ironie, Neutre, Inexploitable. **training samples list** = ['(Tweet Number 1 = Text of tweet 1) (Emotions = Mefiance, Degout, Colere)', '(Tweet Number 2 = Text of tweet 2) (Emotions = Peur, Degout, Colere)', '(Tweet Number 3 = Text of tweet 3) (Emotions = Tristesse, Mefiance, Peur)', '(Tweet Number 4 = Text of tweet 4) (Emotions = Mefiance, Peur, Colere)', '(Tweet Number 5 = Text of tweet 5) (Emotions = Ironie, Surprise, Tristesse)', "(Tweet Number 6 = Text of tweet 6) (Emotions = Surprise, Mefiance, Peur)", '(Tweet Number 7 = Text of tweet 7) (Emotions = Ironie, Mefiance, Degout, Colere)', "(Tweet Number 8 = Text of tweet 8) (Emotions = Neutre, Surprise, Colere)"] The response text should only contain a comma-separated list of up to 4 most likely emotions present in the input French tweet. The emotions should be selected from the emotions list. If you can't find an emotion from the emotions list, simply return the value "Inexploitable"	The prompt closely resembles the one used in zero-shot pseudo-labeling. However, a key distinction lies in the approach: here, we randomly choose a single sample tweet representing each emotion from the training set. Given that a tweet can encompass multiple emotions, it is plausible that emotions beyond the selected one will be present. For instance, consider tweet number 3, primarily chosen due to its association with the emotion *peur* (fear); yet, it concurrently encapsulates the emotions *tristesse* (sadness) and *mefiance* (distrust). A total of 2757 tweets are used for semi-supervised pseudo-labeling after the exclusion of tweets marked as *inexploitable*.
Few-shot	You are a French linguist expert who annotates French tweets with emotions. Pick 4 or less most likely emotions from the emotions list to annotate the input French tweet. As examples use tweet examples and corresponding emotions from the training samples list: **input french tweet** = "C'est l'équivalent d'une marée noire, mais sur terre. Le toit en amiante a été pulvérisé, les gens ont respiré les particules… C'est une vraie catastrophe écologique et sanitaire....." https://t.co/HNNfvQmS3P. **emotions list** = Colere, Degout, Peur, Mefiance, Tristesse, Surprise, Ironie, Neutre, Inexploitable. **training samples list** = ['(Tweet Number 1 = Text of tweet 1) (Emotions = Mefiance, Degout, Colere)', '(Tweet Number 2 = Text of tweet 2) (Emotions = Peur, Degout, Colere)', '(Tweet Number 3 = Text of tweet 3) (Emotions = Tristesse, Mefiance, Peur)', '(Tweet Number 4 = Text of tweet 4) (Emotions = Mefiance, Peur, Colere)', '(Tweet Number 5 = Text of tweet 5) (Emotions = Ironie, Surprise, Tristesse)', "(Tweet Number 6 = Text of tweet 6) (Emotions = Surprise, Mefiance, Peur)", '(Tweet Number 7 = Text of tweet 7) (Emotions = Ironie, Mefiance, Degout, Colere)', "(Tweet Number 8 = Text of tweet 8) (Emotions = Neutre, Surprise, Colere), '(Tweet Number 9 = Text of tweet 9) (Emotions = Mefiance, Degout, Colere)', '(Tweet Number 10 = Text of tweet 10) (Emotions = Peur, Degout, Colere)', '(Tweet Number 11 = Text of tweet 11) (Emotions = Tristesse, Mefiance, Peur)', '(Tweet Number 12 = Text of tweet 12) (Emotions = Mefiance, Peur, Colere)', '(Tweet Number 13 = Text of tweet 13) (Emotions = Ironie, Surprise, Tristesse)', "(Tweet Number 14 = Text of tweet 14) (Emotions = Surprise, Mefiance, Peur)", '(Tweet Number 15 = Text of tweet 15) (Emotions = Ironie, Mefiance, Degout, Colere)', "(Tweet Number 16 = Text of tweet 16) (Emotions = Neutre, Surprise, Colere)"] The response text should only contain a comma-separated list of up to 4 most likely emotions present in the input French tweet. The emotions should be selected from the emotions list. If you can't find an emotion from the emotions list, simply return the value "Inexploitable"	The prompt closely mirrors the structure employed in one-shot pseudo-labeling. However, the prompt incorporates two distinct tweet samples per emotion label, yielding a cumulative count of 16 tweet samples across all emotions. The remaining content of the prompt remains entirely consistent with the format employed in one-shot pseudo-labeling. The exclusion of tweets labelled as *inexploitable* leads to a remaining count of 2775 tweets.

Another interesting result observed in Table 3 is that fully supervised learning models perform better than standard pseudo-labeling models. One possible reason could be that the training dataset used for standard pseudo-labeling in this study is too small. The predicted pseudo-labels could introduce noise during the semi-supervised training step, resulting in reduced performance compared to fully supervised training. This observation aligns with existing research that claims standard pseudo-labeling with small datasets may lead to poor results [80]. Furthermore, it provides support for the foundation of our proposed research, where we employ large language models instead of the standard pseudo-labeling approach.

Table 4 depicts the F1-scores of individual emotions achieved via CamemBERT-Large with Chat-GPT few-shot pseudo-labeling. The findings indicate that, with the exception of *surprise*, the classification performance of all other individual emotions demonstrates improvement compared to fully supervised training. It is noteworthy that the relatively lower performance of *surprise* may be attributed to its reduced representation in the pseudo-labelled annotations as depicted in Figure 3. This finding illustrates that, similar to other deep learning models, there are instances where the pseudo-labels derived from Chat-GPT annotations might not yield optimal performance. This is specially notable for the pseudo-labels that have lower representation. As previously discussed, the underlying cause for the reduced occurrence of a specific label within Chat-GPT annotations could be attributed to a combination of factors, including the characteristics of data used to train Chat-GPT model from scratch, the annotation process, and the specific prompts employed during annotation.

VI. CONCLUSION AND FUTURE WORK

Emotion classification poses a formidable challenge due to the inherent subjectivity of emotions expression. This complexity is further amplified in the realm of multi-label emotion classification, where the potential for conveying multiple emotions within a single textual unit exponentially magnifies subjectivity. Furthermore, the presence of underrepresented emotions within specific contextual scenarios leads to imbalanced datasets. Successful machine learning models for multi-label emotion necessitate a substantial dataset that comprehensively encapsulates the spectrum of human emotional expression. However, the manual annotation of such a dataset requires considerable human supervision, entailing a laborious and costly process.

This study introduces a novel semi-supervised pseudo-labeling approach to multi-label emotion classification, leveraging the capabilities of large language models. We employ Chat-GPT to assign multiple emotion labels to unannotated French tweets in our dataset. Our methodology encompasses zero-shot, one-shot, and few-shot techniques to generate pseudo-labels through Chat-GPT. These pseudo-labels are then merged with manually annotated tweets, facilitating the training of multi-label emotion classification models. Our experimental results substantiate the superiority of our proposed pseudo-labeling-driven semi-supervised learning approach over baseline and state of the art methods. Notably, our approach exhibits enhanced performance on emotion labels that are less frequently represented.

The results further demonstrate that both one-shot and few-shot Chat-GPT annotation techniques return better results compared to baseline and the state of the art methods. Nevertheless, the difference in performance between the one-shot and few-shot learning is minimal. Consequently, we recommend adopting the one-shot approach to minimize the usage of Chat-GPT tokens, thus reducing the annotation cost. On the other hand, zero-shot pseudo-labeling is not recommended due to its inferior results compared to manual annotations.

In addition to proposing a semi-supervised pseudo-labeling approach, we contribute a uniquely tailored French language dataset of tweets annotated with multiple emotions during industrial accidents. This dataset is the first of its kind in French language research for text emotion classification and serves as the foundation for presenting benchmark results in the domain of multi-label emotion classification of French tweets.

While our proposed approach attains state-of-the-art results, several future research directions follow from this study. Firstly, the efficacy of our approach, as demonstrated with Chat-GPT, should be evaluated in comparison with other advanced large language models to ascertain potential performance enhancements. Secondly, the utilization of a more extensive number of tweets per emotion in few-shot training warrants investigation to determine whether it yields improved model performance. Lastly, the introduction of more innovative prompts for Chat-GPT holds the potential to yield more refined annotation outcomes.

APPENDIX
PROMPTS USED FOR PSEUDO LABELING WITH CHAT-GPT

See Table 5.

REFERENCES

[1] J. Brynielsson, F. Johansson, C. Jonsson, and A. Westling, "Emotion classification of social media posts for estimating people's reactions to communicated alert messages during crises," *Secur. Informat.*, vol. 3, no. 1, pp. 1–11, Dec. 2014.

[2] E. Kim and R. Klinger, "A survey on sentiment and emotion analysis for computational literary studies," 2018, *arXiv:1808.03137*.

[3] M. Jabreel and A. Moreno, "A deep learning-based approach for multi-label emotion classification in tweets," *Appl. Sci.*, vol. 9, no. 6, p. 1123, Mar. 2019.

[4] I. Ameer, N. Ashraf, G. Sidorov, and H. Gómez Adorno, "Multi-label emotion classification using content-based features in Twitter," *Computación Sistemas*, vol. 24, no. 3, pp. 1159–1164, Sep. 2020.

[5] K. A. Lindquist, J. K. MacCormack, and H. Shablack, "The role of language in emotion: Predictions from psychological constructionism," *Frontiers Psychol.*, vol. 6, p. 444, Apr. 2015.

[6] E. R. Bailey, S. C. Matz, W. Youyou, and S. S. Iyengar, "Authentic self-expression on social media is associated with greater subjective well-being," *Nature Commun.*, vol. 11, no. 1, p. 4889, Oct. 2020.

[7] B. Vlasenko and A. Wendemuth, "Annotators' agreement and spontaneous emotion classification performance," in *Proc. Interspeech*, 2015, pp. 1546–1550.

[8] H. Hayat, C. Ventura, and A. Lapedriza, "Modeling subjective affect annotations with multi-task learning," *Sensors*, vol. 22, no. 14, p. 5245, Jul. 2022.

[9] J. M. Duarte and L. Berton, "A review of semi-supervised learning for text classification," *Artif. Intell. Rev.*, vol. 56, no. 9, pp. 9401–9469, Sep. 2023.

[10] L. Yang, B. Huang, S. Guo, Y. Lin, and T. Zhao, "A small-sample text classification model based on pseudo-label fusion clustering algorithm," *Appl. Sci.*, vol. 13, no. 8, p. 4716, Apr. 2023.

[11] F. Gilardi, M. Alizadeh, and M. Kubli, "ChatGPT outperforms crowd-workers for text-annotation tasks," 2023, *arXiv:2303.15056*.

[12] I. Ameer, N. Bölücü, M. H. F. Siddiqui, B. Can, G. Sidorov, and A. Gelbukh, "Multi-label emotion classification in texts using transfer learning," *Expert Syst. Appl.*, vol. 213, Jan. 2023, Art. no. 118534.

[13] M. Kutlu, T. McDonnell, M. Lease, and T. Elsayed, "Annotator rationales for labeling tasks in crowdsourcing," *J. Artif. Intell. Res.*, vol. 69, pp. 143–189, Sep. 2020.

[14] *Reuters Over 5,250 Tonnes of Chemicals Burned in Rouen, France Industrial Fire*. Accessed: May 13, 2023. [Online]. Available: https://tinyurl.com/58tny7rx,2019

[15] T. Dopierre, C. Gravier, J. Subercaze, and W. Logerais, "Few-shot pseudo-labeling for intent detection," in *Proc. 28th Int. Conf. Comput. Linguistics*, 2020, pp. 4993–5003. [Online]. Available: https://aclanthology.org/2020.coling-main.438

[16] N. Lin, S. Fu, X. Lin, and L. Wang, "Multi-label emotion classification based on adversarial multi-task learning," *Inf. Process. Manage.*, vol. 59, no. 6, Nov. 2022, Art. no. 103097.

[17] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 649–657.

[18] Y. Gao, Y. Chen, J. Wang, and H. Lu, "Reading scene text with attention convolutional sequence modeling," 2017, *arXiv:1709.04303*.

[19] I. Ameer, N. Bölücü, G. Sidorov, and B. Can, "Emotion classification in texts over graph neural networks: Semantic representation is better than syntactic," *IEEE Access*, vol. 11, pp. 56921–56934, 2023.

[20] H.-D. Le, G.-S. Lee, S.-H. Kim, S. Kim, and H.-J. Yang, "Multi-label multimodal emotion recognition with transformer-based fusion and emotion-level representation learning," *IEEE Access*, vol. 11, pp. 14742–14751, 2023.

[21] S. M. Lakew, M. Cettolo, and M. Federico, "A comparison of transformer and recurrent neural networks on multilingual neural machine translation," 2018, *arXiv:1806.06957*.

[22] S. Tabinda Kokab, S. Asghar, and S. Naz, "Transformer-based deep learning models for the sentiment analysis of social media data," *Array*, vol. 14, Jul. 2022, Art. no. 100157.

[23] G. Soyalp, A. Alar, K. Ozkanli, and B. Yildiz, "Improving text classification with transformer," in *Proc. 6th Int. Conf. Comput. Sci. Eng. (UBMK)*, Sep. 2021, pp. 707–712.

[24] L. Alzubaidi, J. Bai, A. Al-Sabaawi, J. Santamaría, A. S. Albahri, B. S. N. Al-dabbagh, M. A. Fadhel, M. Manoufali, J. Zhang, A. H. Al-Timemy, Y. Duan, A. Abdullah, L. Farhan, Y. Lu, A. Gupta, F. Albu, A. Abbosh, and Y. Gu, "A survey on deep learning tools dealing with data scarcity: Definitions, challenges, solutions, tips, and applications," *J. Big Data*, vol. 10, no. 1, p. 46, Apr. 2023.

[25] C. Yang, Y.-M. Cheung, J. Ding, and K. C. Tan, "Concept drift-tolerant transfer learning in dynamic environments," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 8, pp. 3857–3871, Aug. 2022.

[26] Y. Ouali, C. Hudelot, and M. Tami, "An overview of deep semi-supervised learning," 2020, *arXiv:2006.05278*.

[27] J. Zhang, X. Wang, D. Zhang, and D.-J. Lee, "Semi-supervised group emotion recognition based on contrastive learning," *Electronics*, vol. 11, no. 23, p. 3990, Dec. 2022.

[28] L. Kang, J. Liu, L. Liu, Z. Zhou, and D. Ye, "Semi-supervised emotion recognition in textual conversation via a context-augmented auxiliary training task," *Inf. Process. Manage.*, vol. 58, no. 6, Nov. 2021, Art. no. 102717.

[29] M. Jiang, Y. Liang, X. Feng, X. Fan, Z. Pei, Y. Xue, and R. Guan, "Text classification based on deep belief network and softmax regression," *Neural Comput. Appl.*, vol. 29, no. 1, pp. 61–70, Jan. 2018.

[30] Y. Pan, Z. Chen, Y. Suzuki, F. Fukumoto, and H. Nishizaki, "Sentiment analysis using semi-supervised learning with few labeled data," in *Proc. Int. Conf. Cyberworlds (CW)*, Sep. 2020, pp. 231–234.

[31] Y. Cheng, K. Qian, Y. Wang, and D. Zhao, "Missing multi-label learning with non-equilibrium based on classification margin," *Appl. Soft Comput.*, vol. 86, Jan. 2020, Art. no. 105924.

[32] A. Krishnamoorthy, A. K. Patil, N. Vasudevan, and V. Pathari, "News article classification with clustering using semi-supervised learning," in *Proc. Int. Conf. Adv. Comput., Commun. Informat. (ICACCI)*, Sep. 2018, pp. 86–91.

[33] L. A. Vilhagra, E. R. Fernandes, and B. M. Nogueira, "TextCSN: A semi-supervised approach for text clustering using pairwise constraints and convolutional Siamese network," in *Proc. 35th Annu. ACM Symp. Appl. Comput.*, Mar. 2020, pp. 1135–1142.

[34] L. H. X. Ng and K. M. Carley, "'The coronavirus is a bioweapon': Classifying coronavirus stories on fact-checking sites," *Comput. Math. Org. Theory*, vol. 27, no. 2, pp. 179–194, Jun. 2021.

[35] Y. Qiu, X. Gong, Z. Ma, and X. Chen, "MixLab: An informative semi-supervised method for multi-label classification," in *Proc. CCF Int. Conf. Natural Lang. Process. Chin. Comput.*, Zhengzhou, China, Oct. 2020, pp. 506–518.

[36] A. H. Li and A. Sethy, "Semi-supervised learning for text classification by layer partitioning," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 6164–6168.

[37] R. Gupta, S. Sahu, C. Espy-Wilson, and S. Narayanan, "Semi-supervised and transfer learning approaches for low resource sentiment classification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 5109–5113.

[38] Y. Grandvalet and Y. Bengio, "Semi-supervised learning by entropy minimization," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 17, 2004, pp. 529–536.

[39] Y. Fan, A. Kukleva, D. Dai, and B. Schiele, "Revisiting consistency regularization for semi-supervised learning," *Int. J. Comput. Vis.*, vol. 131, no. 3, pp. 626–643, Mar. 2023.

[40] D. Kingma, S. Mohamed, D. J. Rezende, and M. Welling, "Semi-supervised learning with deep generative models," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 3581–3589.

[41] S. Park, J. Lee, and K. Kim, "Semi-supervised distributed representations of documents for sentiment analysis," *Neural Netw.*, vol. 119, pp. 139–150, Nov. 2019.

[42] S. Shehnepoor, R. Togneri, W. Liu, and M. Bennamoun, "ScoreGAN: A fraud review detector based on regulated GAN with data augmentation," *IEEE Trans. Inf. Forensics Security*, vol. 17, pp. 280–291, 2022.

[43] X. Yang, Z. Song, I. King, and Z. Xu, "A survey on deep semi-supervised learning," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 9, pp. 8934–8954, Sep. 2023.

[44] D. Lee, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Proc. Workshop Challenges Represent. Learn.*, vol. 3, 2013, p. 896.

[45] P. Cascante-Bonilla, F. Tan, Y. Qi, and V. Ordonez, "Curriculum labeling: Revisiting pseudo-labeling for semi-supervised learning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, 2021, pp. 6912–6920.

[46] A. Oliver, A. Odena, C. A. Raffel, E. D. Cubuk, and I. Goodfellow, "Realistic evaluation of semi-supervised learning algorithms," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–15.

[47] E. Arazo, D. Ortego, P. Albert, N. E. O'Connor, and K. McGuinness, "Pseudo-labeling and confirmation bias in deep semi-supervised learning," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2020, pp. 1–8.

[48] N. Widmann and S. Verberne, "Graph-based semi-supervised learning for text classification," in *Proc. ACM SIGIR Int. Conf. Theory Inf. Retr.*, Oct. 2017, pp. 59–66.

[49] M. C. de Souza, B. M. Nogueira, R. G. Rossi, R. M. Marcacini, and S. O. Rezende, "A heterogeneous network-based positive and unlabelled learning approach to detect fake news," in *Proc. Brazilian Conf. Intell. Syst.*, 2021, pp. 3–18.

[50] J. Bose and S. Mukherjee, "Semi-supervised method using Gaussian random fields for boilerplate removal in web browsers," in *Proc. IEEE 16th India Council Int. Conf. (INDICON)*, Dec. 2019, pp. 1–4.

[51] H. Zhao, J. Xie, and H. Wang, "Graph convolutional network based on multi-head pooling for short text classification," *IEEE Access*, vol. 10, pp. 11947–11956, 2022.

[52] D.-H. Zhu, X.-Y. Dai, and J.-J. Chen, "Pre-train and learn: Preserving global information for graph neural networks," *J. Comput. Sci. Technol.*, vol. 36, no. 6, pp. 1420–1430, Dec. 2021.

[53] W. Ju, J. Yang, M. Qu, W. Song, J. Shen, and M. Zhang, "KGNN: Harnessing kernel-based networks for semi-supervised graph classification," in *Proc. 15th ACM Int. Conf. Web Search Data Mining*, Feb. 2022, pp. 421–429.

[54] C. Li, X. Li, and J. Ouyang, "Semi-supervised text classification with balanced deep representation distributions," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process.*, 2021, pp. 5044–5053.

[55] H. Hwang and Y. Lee, "Semi-supervised learning based on auto-generated lexicon using XAI in sentiment analysis," in *Proc. Conf. Recent Adv. Natural Lang. Process.-Deep Learn. Natural Lang. Process. Methods Appl.*, 2021, pp. 593–600.

[56] H. Zou and Z. Wang, "A semi-supervised short text sentiment classification method based on improved BERT model from unlabelled data," *J. Big Data*, vol. 10, no. 1, pp. 1–19, Mar. 2023.

[57] Y. Chong, Y. Ding, Q. Yan, and S. Pan, "Graph-based semi-supervised learning: A review," *Neurocomputing*, vol. 408, pp. 216–230, Sep. 2020.

[58] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. A. Raffel, "MixMatch: A holistic approach to semi-supervised learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–14.

[59] D. Mekala, C. Dong, and J. Shang, "LOPS: Learning order inspired pseudo-label selection for weakly supervised text classification," 2022, *arXiv:2205.12528*.

[60] Z. Xinxi, "Single task fine-tune BERT for text classification," in *Proc. 2nd Int. Conf. Comput. Vis., Image, Deep Learn.*, Kunming, China, Oct. 2021, pp. 194–206.

[61] D. Jimenez, O. J. Gambino, and H. Calvo, "Pseudo-labeling improves news identification and categorization with few annotated data," *Computación Sistemas*, vol. 26, no. 1, pp. 183–193, Mar. 2022.

[62] A. Anaby-Tavor, B. Carmeli, E. Goldbraich, A. Kantor, G. Kour, S. Shlomov, N. Tepper, and N. Zwerdling, "Do not have enough data? Deep learning to the rescue!" in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, 2020, pp. 7383–7390.

[63] R. Chandra and A. Krishna, "COVID-19 sentiment analysis via deep learning during the rise of novel cases," *PLoS ONE*, vol. 16, no. 8, Aug. 2021, Art. no. e0255615.

[64] M. Y. Kabir and S. Madria, "EMOCOV: Machine learning for emotion detection, analysis and visualization using COVID-19 tweets," *Online Social Netw. Media*, vol. 23, May 2021, Art. no. 100135.

[65] (2019). *The Local France French Chemical Factory Explosion: What We Know So Far*. Accessed: May 13, 2023. [Online]. Available: https://tinyurl.com/29ft4xva

[66] (2019). *France 24 French Authorities Probe Lubrizol Factory Fire in Normandy*. Accessed: May 13, 2023. [Online]. Available: https://tinyurl.com/kdjc96k7

[67] (2019). *The New York Times Fire at French Chemical Plant Provokes Worry Across the Channel*. Accessed: May 13, 2023. [Online]. Available: https://tinyurl.com/3565bjyv

[68] R. Plutchik, *A Psychoevolutionary Theory of Emotions*. Newbury Park, CA, USA: Sage, 1982.

[69] P. Nandwani and R. Verma, "A review on sentiment analysis and emotion detection from text," *Social Netw. Anal. Mining*, vol. 11, no. 1, p. 81, Dec. 2021.

[70] J. Mekki, G. Lecorvé, D. Battistelli, and N. Béchet, "TREMoLo-tweets: A multi-label corpus of French tweets for language register characterization," in *Proc. Int. Conf. Recent Adv. Natural Lang. Process.*, 2021, pp. 950–958.

[71] K. Krippendorff, "Computing Krippendorff's alpha-reliability," Annenberg School Commun., Univ. Pennsylvania, Philadelphia, PA, USA, Tech. Rep., 2011.

[72] J. Fleiss, B. Levin, and M. Paik, *Statistical Methods for Rates and Proportions*. Hoboken, NJ, USA: Wiley, 2013.

[73] K. Krippendorff, *Content Analysis: An Introduction to Its Methodology*. Newbury Park, CA, USA: Sage, 2018.

[74] M. L. McHugh, "Interrater reliability: The Kappa statistic," *Biochemia Medica*, vol. 22, pp. 276–282, Nov. 2012.

[75] L. Martin, B. Müller, P. J. O. Suárez, Y. Dupont, L. Romary, É. de la Clergerie, D. Seddah, and B. Sagot, "CamemBERT: A tasty French language model," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 7203–7219.

[76] H. Le, L. Vial, J. Frej, V. Segonne, M. Coavoux, B. Lecouteux, A. Allauzen, B. Crabbé, and L. Besacier, "FlauBERT: Unsupervised language model pre-training for French," in *Proc. 12th Lang. Resour. Eval. Conf.*, 2020, pp. 2479–2490. [Online]. Available: https://aclanthology.org/2020.lrec-1.302

[77] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.

[78] R. Sauvayre, J. Vernier, and C. Chauvière, "An analysis of French-language tweets about COVID-19 vaccines: Supervised learning approach," *JMIR Med. Informat.*, vol. 10, no. 5, May 2022, Art. no. e37831.

[79] G. I. Winata and M. L. Khodra, "Handling imbalanced dataset in multi-label text categorization using bagging and adaptive boosting," in *Proc. Int. Conf. Electr. Eng. Informat. (ICEEI)*, Aug. 2015, pp. 500–505.

[80] H. Wu, X. Li, Y. Lin, and K.-T. Cheng, "Compete to win: Enhancing pseudo labels for barely-supervised medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 42, no. 11, pp. 3244–3255, Apr. 2023.

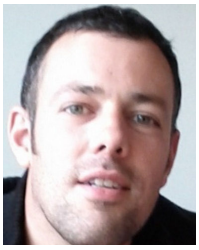
[81] G. Nguyen, S. Chen, T. Jun, and D. Kim, "Explaining how deep neural networks forget by deep visualization," in *Proc. Int. Conf. Pattern Recognit. Workshops*, 2020, pp. 1–16.

[82] J. Lee, R. Tang, and J. Lin, "What would Elsa do? Freezing layers during transformer fine-tuning," 2019, *arXiv:1911.03090*.

USMAN MALIK received the Ph.D. degree from Normandie Université, France, where his research focused on the application of machine learning and deep learning techniques to enhance multi-model human-agent interaction. Currently, he is actively engaged in the CATCH project, where his focus lies in the development of multi-label emotion classification models tailored for French-language tweets.

SIMON BERNARD received the Ph.D. degree in computer science from the University of Rouen Normandy, France, in 2009. He has been an Associate Professor (Maitre de Conféences) with the University of Rouen, Normandy, since 2013. He is also a member of the LITIS Laboratory and Normastic CNRS Research Federation. His research interests include machine learning, ensemble learning, and deep learning, in various learning contexts, such as weakly supervised learning, one-class classification and/or high-dimensional, and low sample size (HDLSS) classification.

ALEXANDRE PAUCHET is currently an Associate Professor with INSA Rouen Normandy, France, and the LITIS Laboratory. He is also in charge of the MIND Team and a Co-Animator of the Data, Learning and Knowledge Axis of the Normastic Federation. He has defended the French ability to supervise researches (HDR) with the University of Rouen, in 2015. His research interests include human-agent interaction, affective computing, and dialogue.



CLÉMENT CHATELAIN is currently an Associate Professor with the Department of Information Systems Engineering, INSA Rouen Normandy, France. His research interests include machine learning applied to handwriting recognition, document image analysis, and medical image analysis. His teaching and research interests include signal processing, deep learning, and pattern recognition. In 2019, he received the French ability to supervise researches from the University of Rouen.

ROMAIN PICOT-CLÉMENTE received the Ph.D. degree in computer science from the University of Burgundy, France, in 2011. He has been the Head of Artificial Intelligence with Saagie, since 2018. He is currently the Co-Director of the Joint Laboratory L-LiSa between LITIS and Saagie. His research interests include deep learning, weakly supervised learning applied to text, and time series data.

JÉRÔME CORTINOVIS received the Ph.D. degree in physico-chemical modeling of the atmosphere from CNRS-Laboratoire d'Aérologie, Toulouse, in 2004. He has been an Innovation and Partnership Engineer with Atmo Normandie, since 2004. He has worked in particular on the development of numerical modelling of air quality and on the implementation of open data. He is currently a Coordinator of Incub'air, Atmo Normandie's Innovation Laboratory, which aims to test and disseminate innovative solutions for air quality (including odor issues).