CrisisCL: a Domain Incremental Learning Benchmark for Crisis Management

Paul Le Van Kiem[3], Romain Meunier[1], Farah Benamara[1,2], Véronique Moriceau[1]
(1) Univ Toulouse, Toulouse INP, CNRS, IRIT, Toulouse, France
(2) IPAL, CNRS-NUS-A*STAR, Singapore
(3) CNRS@CREATE LTD, Singapore
paul.le-van-kiem@cnrsatcreate.sg, {firstname.lastname}@irit.fr

Abstract

This paper proposes **CrisisCL**, a domain incremental learning benchmark for crisis management. Based on previous crisis management protocols, it improves consistency by allowing continual learning (CL) of new crises. A set of experiments have been conducted on multilingual datasets relying on continual learning methods and transformers to improve performance and ensure model generalization. Results reveal that regularization methods are more effective on large, coherent domains, whereas replay strategies struggle under constrained memory. Additional experimental protocols further expose the limitations of current CL methods when generalizing to unforeseen crisis events.

Keywords: Datasets, Continual learning, Crisis management

1. Introduction

Natural disasters and humanitarian crises unfold suddenly generate an urgent need for timely and structured information. Social media platforms, particularly X (ex-Twitter), have emerged as de facto "social sensors" (Alam et al., 2021), providing real-time insights within minutes of an event. Yet these data streams are vast, noisy, and highly dynamic: only a small fraction of posts is relevant to responders, and the language used shifts dramatically across different types of crises, regions, and time. Traditional NLP pipelines are typically trained once using given data and then frozen. This static approach faces two main challenges: concept drift, which involves the evolution of linguistic patterns over time, and task evolution, which pertains to the emergence of new label categories. Fine-tuning, risks catastrophic forgetting or re-training from scratch is often impossible due to constraints such as data scarcity, privacy concerns, or limited time.

Continual Learning (CL) offers a promising alternative by enabling models to learn incrementally without forgetting past knowledge (Ke and Liu, 2022). While extensively studied in computer vision (Van de Ven and Tolias, 2019), CL remains underexplored in NLP—particularly in crisis management settings. Existing studies often focus on single tasks or datasets, and to the best of our knowledge, no standard benchmark currently evaluates CL methods on multilingual, real-world crisis data (Bourgon et al., 2022; Priyanshu et al., 2021).

In this paper, we focus on adapting and evaluating continual learning techniques for multilingual crisis tweet classification. In this context, our contributions are:

- A new benchmark, CrisisCL, for Domain-Incremental Learning (DIL) inspired by state-of the art crisis management *out-of-type* experimental scenario (Kersten et al., 2019; Algiriyage et al., 2021; Bourgon et al., 2022). Unlike Class-Incremental Learning (CIL), which introduces new label categories over time, the DIL setting assumes a fixed label space and models changes in the input distribution — e.g. a model trained on one crisis (e.g., flood) must adapt to another one (e.g., wildfire).
- The implementation and evaluation of several classical continual learning methods on multiple datasets for crisis management and different languages.

In the following, Section 2 summarizes the related work. Sections 3 and 4 respectively present the CrisisCL benchmark as well as the experimental settings. Section 5 presents our results. Finally, we conclude drawing some perspectives for future work.

2. Related Work

2.1. NLP for Crisis Management

In crisis situations, real-time textual information can come from the emergency services (Otal and Canbaz, 2024) but also from social media (Reuter et al., 2018) (e.g., more than one million tweets posted during 2023 Turkey–Syria earthquakes (Toraman et al., 2023)). In order to collect, extract or summarize this information, NLP-based crisis management has become a hot research topic in text classification where posted messages are classified into different categories, named entity recognition

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 10853–10865
11–16 May 2026.

to detect location mention that helps geolocalise information needs (Suwaileh et al., 2023), and event detection (Rajaby et al., 2022).

Text-based classifiers are mainly trained in a supervised way with either traditional feature-based learning algorithms (Li et al., 2018; Kaufhold et al., 2020; Alam et al., 2021) or deep learning architectures (Caragea et al., 2016; Castillo, 2016; Neppalli et al., 2018; Kersten et al., 2019; Kozlowski et al., 2020; Chowdhury et al., 2020; Liu et al., 2021; Wang et al., 2021; Dusart et al., 2021). More recent work use generative AI (Otal and Canbaz, 2024) or combine several modalities to improve urgency classification (Alam et al., 2018).

While de Bruijn et al. (2020) focus on one crisis such as flood, dealing with different crises is a more challenging task. Recent work focus on multiple crises with multi-crises datasets such as TREC-IS (McCreadie et al., 2019), CrisisFACTS (McCreadie and Buntain, 2023) and CrisisTS (Meunier et al., 2025), with three main tasks: (1) relatedness (Is a message useful/relevant for emergency departments?), (2) urgency (Is a message urgent for emergency departments? Possibly what is the urgency degree?), and (3) humanitarian categories (damage reports, warnings, critics, etc.). Overall, results show that humanitarian categories detection is the most challenging task due to the extremely imbalanced nature of crisis datasets.

2.2. Continual Learning

Continual Learning is about ensuring that a model is able to learn new tasks without forgetting older ones (Wang et al., 2024). In traditional training, there are numerous datasets on which the model can train. However, in some scenario, the data can evolve (e.g. new language expression) or occur requiring training on new data without forgetting information from previous training.

2.2.1. Continual Learning Scenarios

Continual learning represents all the methods used to mitigate forgetting from previous training. In order to create a simulation of this condition, several scenarios have been created:

- **Instance-Incremental Learning (IIL)** (Beringer and Hüllermeier, 2007; Zhang et al., 2011): All training samples belong to the same task and arrive in batches so the model learns from each training example as it arrives.
- **Domain-Incremental Learning (DIL)** (Acharya et al., 2020): In this scenario, tasks have the same data label space but different input distributions. Task identities are not required.
- **Task-Incremental Learning (TIL)** (De Lange et al., 2021): Tasks have disjoint data label spaces. Task identities are provided in both training and testing.
- **Class-Incremental Learning (CIL)** (Mai et al., 2022): Such as TIL, tasks have disjoint data label spaces. However task identities are only provided in training.
- **Task-Free Continual Learning (TFCL)** (Aljundi et al., 2019b): Here, tasks also have disjoint data label spaces but task identities are not provided in either training nor testing.
- **Online Continual Learning (OCL)** (Aljundi et al., 2019c): In this scenario, training samples for each task arrive as a one-pass data stream. The tasks have still disjoint data label spaces.
- **Blurred Boundary Continual Learning (BBCL)** (Bang et al., 2022): This scenario is used when there is incertitude in the definition of the multiple tasks. Here, task boundaries are blurred, characterized by distinct but overlapping data label spaces.

2.2.2. Regularization-based Approaches

The previous scenarios help to simulate continual learning so that models do not forget information from previous training. There are three main approaches to prevent forgetting.

Weights Regularization. It consists in adding a penalty term to the loss function in order not to forget the older task (Kozal et al., 2024). For example, the *Elastic Weights Consolidation* (EWC) (Kirkpatrick et al., 2017) method estimates the importance of network weights using the Fisher Information matrix and penalizes deviations from their previous values during training on a new task. The *Synaptic Intelligence* method (SI) (Zenke et al., 2017) is motivated by the observation that parameters contributing significantly to reducing the loss on previous tasks should be preserved. Unlike methods such as EWC which estimate weight importance after training, SI computes weight importance online throughout training, without requiring additional passes or storage of task-specific data. Finally, *Memory Aware Synapses* (Aljundi et al., 2018) is a method used to handle unlabeled data. Therefore, only the model and the different gradients are needed. Moreover, this method also handles online learning as the samples can be easily added over time.

Loss Function Regularization. Other methods try to regularize the loss function. It can be achieved via knowledge distillation or Bayesian inference. For example, Li and Hoiem (2017) proposed a method called *Learning Without Forgetting* which incorporates knowledge distillation to train the network on new tasks while preserving the output probabilities of the previous tasks. One of the most important function regularization method is iCaRL (Rebuffi et al., 2017) (Incremental Classifier and Representation Learning) which addresses class-incremental learning by combining prototype-based classification, exemplar management, and representation learning with distillation.

Replay-Based Approach. This method aims to mitigate catastrophic forgetting by approximating or recovering old data distributions. This is done by replaying previously seen data or their proxies when learning new tasks. *Gradient of Episodic Memory* (GEM)(Lopez-Paz and Ranzato, 2022), a method where each task t has an episodic memory $\mathcal{M}_t$ that stores a subset of examples from that task. The loss over the memory of a past task k is defined as:

$$\ell(f_\theta, \mathcal{M}_k) = \frac{1}{|\mathcal{M}_k|} \sum_{(x_i, y_i) \in \mathcal{M}_k} \ell(f_\theta(x_i, k), y_i).$$

At each training step, GEM minimizes the loss on the current sample (x, t, y), while constraining the loss on previous tasks not to increase:

$$\begin{aligned} \min_{\theta} \quad & \ell(f_\theta(x, t), y) \\ \text{subject to} \quad & \ell(f_\theta, \mathcal{M}_k) \leq \ell(f_\theta^{t-1}, \mathcal{M}_k), \quad \forall k < t. \end{aligned}$$

Unlike methods that rely on storing and replaying actual data, *Deep Generative Replay* (DGR) framework (Shin et al., 2017) allows sequential learning on multiple tasks by generating and rehearsing fake data that mimics former training examples. *FearNet* (Kemker and Kanan, 2017) follows the proposition of DGR with a dual-memory incremental learning framework.

2.3. Continual Learning in NLP

While early research on CL primarily focused on image classification tasks in Computer Vision (Li and Hoiem, 2017), with the introduction of generative AI such as GPT-4 (OpenAI et al., 2024), CL has become a very important part of NLP. Indeed, such models fundamentally depend on large-scale generative pre-training and aligning with human insights. Nevertheless, this method is intrinsically dynamic given the constant evolution of languages, data distributions, and user needs. Finally, it is difficult to foresee every potential future scenario in advance, highlighting the necessity for models to exhibit CL abilities akin to humans to effectively handle real-world situations (Zhou, 2022).

Some previous work on CL in NLP include LAMOL (Sun et al., 2019), which proposes a pseudo-replay-based approach learning downstream tasks. It also learns to generate training data for downstream tasks simultaneously. CL is also used for example in machine translation tasks (Berard, 2021; Chuang et al., 2020). More recently, Satapara and Srijith (2024) proposed a benchmark that combines incremental task learning and incremental multilingual learning.

Most approaches apply CL to address model portability across tasks and languages, overlooking how models can deal with domain shift. Recently, Jain et al. (2025) proposed a domain incremental learning approach to detect the dynamic evolution of specific vocabulary across various domains such as business, sports and technology. Besides an analysis of vocabulary change, they do not experimentally show how this change may impact model performances on a specific domain. Our work goes one step further by proposing the first benchmark for domain incremental learning for crisis management and a set of CL scenarios to measure the portability of urgency detection models when dealing with unseen crises.

3. CrisisCL: A Domain Incremental Learning Benchmark

In crisis management, one of the main challenge is to evaluate models in an *out-of-type* protocol (Kersten et al., 2019; Algiriyage et al., 2021; Bourgon et al., 2022): it consists in the average of n runs, each run with $n-1$ crisis types for training and the remaining crisis type for testing. It aims to evaluate if a model can deal with new types of crisis. This task is especially challenging since each type of crisis can have its own lexicon and even its own distribution (e.g. a sudden crisis such as an earthquake will have less tweets about ADVICE OR WARNING than a hurricane for which its impact can be predicted several days before) (Meunier et al., 2025). Therefore, in a real life scenario, if the model faces a new type of crisis, the performance can drop significantly, even on previous crises (for example, Meunier et al. (2025) observed a drop in F1-score between 3 point and 7 points).

This scenario fits in a Domain-Incremental Learning (DIL) setting, where the data is partitioned into distinct domains and each domain corresponds to a specific crisis type (e.g., flood, earthquake, wildfire...). Each domain contains its own training and test set. Unlike Task-Incremental Learning, the output space remains unchanged across all domains. However, the input distribution shifts between domains due to the specific vocabulary,

linguistic patterns, and contextual cues inherent to each crisis type. The model must thus learn to generalize across domains without explicit task identifiers and without access to previous domain data during training. In the following sections, we present the different datasets used to design an DIL framework.

3.1. Datasets

We built our continual learning benchmark for crisis management on different existing datasets in French and English:

- **Kozlowski** (Kozlowski et al., 2020) : It is the largest corpus of French tweets annotated for crisis and augmented later on by Bourgon et al. (2022). It is composed of 7 types of crisis (Fire, Flood, Storm, Hurricane, Collapse, Explosion, Attack) with several crises which occurred in France such as Notre-Dame fire or flood in the Aude region. It contains tweets, collected 24h before, during (48h) and up to 72h after the crisis, manually annotated for three urgency categories as well as 6 intent to act categories (similar to humanitarian categories): (1) Urgent that applies to messages mentioning Human/Material damages as well as security instructions (Advice-Warning) to limit these damages during crisis events, (2) Not urgent that groups Support messages to the victims, Critics or any Other messages that do not have an immediate impact but contribute in raising situational awareness, and finally (3) Not useful for messages that are not related to the targeted crisis.
- **HumAID** (Alam et al., 2021): It is one of the largest publicly available Twitter corpus annotated for humanitarian information extraction. Starting from a pool of 24M tweets gathered during 19 major disasters that occurred between 2016 and 2019 including hurricanes, earthquakes, wildfires and floods. A stratified sampling step then yielded 77K English tweets that are likely to originate from the disaster-hit areas. Each tweet is manually labeled via Amazon Mechanical Turk with one of 11 humanitarian categories reflecting critical information needs (e.g., Caution&Advice, Requests or Urgent Needs, Infrastructure Damage, Sympathy&Support, etc.).

In order to have parallel French–English datasets and to better compare the results obtained on both datasets, we created the Urgent, Not Urgent and Not Useful labels for HumAID by mapping the French and English humanitarian labels. Detailed distributions of Kozlowski and HumAID datasets are presented in Table 1 and Table 2 respectively.

3.2. Consistent Sampling of Data

3.2.1. Training et testing splits

As both datasets have different size and distribution, we used different sampling methods to build the different training and testing sets:

- **Kozlowski splits:** Because the French dataset is smaller, we built three disjoint test sets for finer-grained evaluation. Each test set isolates exactly one event per crisis type, while the remaining tweets form the shared training pool.[1] Due to a lack of data, Attack and Explosion were removed as they hardly contain 100 annotated tweets.
  - All crisis: Flood (Aude, Corsica, "Other"), Storm (Corsica, Beryl-Guadeloupe, Bruno, Egon, Eleanor, Bregitta, Susanna, Ulrika) Hurricane (Irma, Harvey, Fire (Notre Dame, Lande) Terrorist attack (Trèbes), Collapse (Marseilles, Lille), Explosion (Sanary, Lubrizol).
  - Test Set 1: Flood in Aude, Storm in Beryl-Guadeloupe, Collapse in Lille and Hurricane Harvey.
  - Test Set 2: Flood in Corsica, Storm in Corsica, Collapse in Marseilles and Hurricane Harvey
  - Test Set 3: Flood "Other", Storm Egon, Collapse in Lille and Hurricane Irma
- **HumAID split:** As the HumAID dataset is much larger and in order to get comparable results, we decided to sample the training set to contain at most 2,000 tweets per crisis type, chosen from one representative event for that type; the rest constitutes the test set. By testing different random sampling, we noticed that the results do not change significantly, therefore we decided to keep only one sampling:
  - Train Set: 2016 Wildfire in Canada, 2018 Hurricane in Florence, 2016 Earthquake in Kaikoura, and 2019 Flood in Mid western US
  - Test set: 2016 Ecuador Earthquake, 2016 Italy Earthquake, 2016 Hurricane Matthew, 2017 Sri Lanka Floods, 2017 Hurricane Harvey, 2017 Hurricane Irma, 2017 Hurricane Maria, 2017 Mexico Earthquake, 2018 Maryland Floods, 2018 Greece Wildfires, 2018 Kerala Floods, 2018 California Wildfires, 2019 Cyclone Idai, 2019 Hurricane Dorian, 2019 Pakistan Earthquake

[1] For a test set i, all crises that are not in test set belong to the train set

Crisis (# events / # tweets)	Urgent			Not Urgent			Not Useful
	Adv_Warn	Hmn-Dmg	Mat-Dmg	Critics	Support	Other	
Not Sudden (13 crises / 11,513 tweets)							Not Crisis
Flood (3 / 4,190)	493	108	280	58	244	481	2,526
Storm (8 / 5,762)	716	52	142	13	22	147	4,670
Hurricane (2 / 2,160)	199	57	57	29	200	200	1,418
Sudden (7 crises / 3,855 tweets)							Not Crisis
Fire (2 / 2,443)	50	23	93	166	340	379	1,392
Attack (1 / 45)	0	3	0	2	38	0	2
Collapse (2 / 1,269)	11	63	38	51	23	136	947
Explosion (2 / 56)	0	1	3	0	52	0	0
Total (20 / 15,368)	1,469	307	613	319	1,119	1,343	10,955

Table 1: Distribution of textual data in Kozlowski dataset.

Crisis (# events / # tweets)	Urgent					Not Urgent			Not Useful
	Adv_Warn	Evac	Mat-Dmg	Hmn-Dmg	Urg_Needs	Other	Help	Support	
Not Sudden (7 crises / 33,233 tweets)									Not Crisis
Flood (3 / 10,424)	246	56	453	409	670	1,284	5,401	1,040	865
Hurricane (4 / 22,809)	2,935	1,608	2,740	883	997	3,976	4,768	2,622	2,280
Sudden (4 crises / 18,371 tweets)									Not Crisis
Fire (2 / 9,439)	245	748	673	1,946	99	1,349	2,349	633	1,397
Earthquake (2 / 8,932)	629	87	728	1,489	225	707	2,049	2,520	498
Total (14 / 51,604)	4,055	2,499	4,594	4,727	1,991	7,316	14,567	6,815	5,040

Table 2: Distribution of textual data in HumAID dataset.

After this step, both datasets provide: (i) a continual learning training set composed of 4 crisis types, (ii) a unified three-task classification set of labels, and (iii) train/test partitions tailored to our continual-learning protocols.

3.2.2. Domain order

As we can see in Tables 1 and 2, crisis types are imbalanced. Indeed, in Kozlowski dataset, 8.25% of the tweets belongs to crisis type *Collapse* and 37.49% to *Storm*. This means that regarding which crisis type is in the train set, we can have a huge difference in training size. Moreover, in HumAID, even if we flatten the number of tweets per crisis, the distribution of the labels between crises is really different with 3.8 % of hurricane tweets that are labeled as Human Damage while in fire, there is 20.61% of Human Damage. Therefore, each domain (or crisis type) has different quality and brings different information. Thus, the final performance of the model may change depending on the crisis order during training. Therefore, we propose a protocol where all the crisis types order are considered. Let $C = \{c1, c2, c3, c4\}$ be the four crisis types of the training set (sorted alphabetically). For each train–test split we create four rotating testing order:

- Order 1: c1 →c2 →c3 →c4,
- Order 2: c2 →c3 →c4 →c1,
- Order 3: c3 →c4 →c1 →c2,
- Order 4: c4 →c1 →c2 →c3,

so that every crisis type occupies each temporal position exactly once, mitigating order bias. For HumAID, as there is a single split, every CL method is trained on the four orders (four CL runs), and this process was repeated 4 times in order to reduce randomness, giving 4 × 4 = 16 CL runs. For Kozlowski dataset, as we already use three independent splits to reduce randomness, we only do the whole process 3 times giving 3 × 4 × 3 = 36 CL runs per method. The results are the average over these runs.

4. Experimental Settings

The experimental settings reflect practical deployment challenges in crisis management, where a system trained on past crises must be adapted to new, unseen events without forgetting prior knowledge. After each training step on domain D_i, we evaluate the model on all previously seen domains (D_1 to D_i) using the F1-Macro score for each domain (see Algorithm 1).

4.1. Continual Learning Protocol (CLP)

In order to evaluate our benchmark, we rely on several state of the art continual learning methods, focusing on those widely used for domain incremental learning for NLP (Michieli and Ozay, 2024; Wang et al., 2022):

- Weights Regularization approaches:
  - **Elastic Weight Consolidation (EWC)**: Introduced by Kirkpatrick et al. (2017),

Algorithm 1: CRISISCL PROCEDURE

Input: Crisis datasets $D_1, ..., D_D$, each corresponding to a domain

Output: Mean F1-Macro scores $F1_{i,j}$ with $1 \leq i \leq D, 1 \leq j \leq i$

```
for i = 1 to D do
    Train model θ on D_i^train
    for j = 1 to i do
        Evaluate model on D_j^test and
        compute F1-macro F1_{i,j}
```

EWC is a regularization-based approach that avoids catastrophic forgetting by limiting the learning of critical parameters for previous domains. This method uses the Fisher information matrix to calculate the importance of each parameter.

  - **SYNAPTIC INTELLIGENCE (SI)**: Proposed by Zenke et al. (2017) and motivated by the observation that parameters contributing significantly to reducing the loss on previous tasks should be preserved, SI computes importance online throughout training, without requiring additional passes or storage of task-specific data.
  - **MEMORY-AWARE SYNAPSES (MAS)** (Aljundi et al., 2018) is used to handle unlabeled data. It computes the importance of the parameters of a neural network in an unsupervised and online manner.

- Replay-based approaches:
  - **AVERAGE GEM (A-GEM)**: Introduced by Chaudhry et al. (2019), it is an optimized version of Gradient Episodic Memory. It tries to ensure that at every training step, the average episodic memory loss over the previous tasks does not increase.
  - **NAIVE EXPERIENCE REPLAY (NER)** (Aljundi et al., 2019a): this replay-based method stores a subset of representative examples from previous tasks and use them to augment the data in the incoming batch.

In addition, we consider as baselines: (1) the **Vanilla** Sequential Finetuning (i.e. no Continual Learning method) which offers a naive lower bound, as it is prone to Catastrophic Forgetting; and (2) the **Cumulative** baseline which is the best scenario model that stores all the data every time a new dataset is available (it is considered as the best case scenario and is used as an upper bound).

4.2. Models

We evaluate the following models on the previous benchmarks and continual learning protocols. The models were trained in a multitask configuration with one classification head per task (utility, urgency and humanitarian):

- **FlauBERT**$_{FineTuned\_3Tasks}$: It is a FlauBERT model (Le et al., 2020) that has been fine-tuned on 358,834 unlabelled tweets posted during crises. It is reported to be the best performing model for intent classification on the French KOSLOWSKI dataset (Meunier et al., 2023). We use it on the French dataset.

- **RoBERTa**$_{3Tasks}$ This is a baseline model for the English dataset as it is reported to be efficient for crisis management in English (Koshy and Elango, 2023; Rocca et al., 2023; Madichetty et al., 2023).

All models are optimized with Adam ($\beta 1 = 0.9$, $\beta 2 = 0.999$). Algorithm-specific hyper-parameters (memory size, regularization strength, etc.) are summarized in Table 3 and were tuned on a training split for each dataset with a classic grid search.

4.3. Computing Infrastructure

In order to improve the reproducibility of the experiments, and as some methods of continual learning such as A-GEM are sensitive to memory sizes, we describe here the computing infrastructure we used: Google Colab with an A100 GPU with 40GB of RAM.

4.4. Metrics

In continual learning, several metrics are commonly used to evaluate model performance over a sequence of domains. Let $a_{i,j}$ be the F1-score after training on domain i, testing on domain j and T the total number of domains. The following metrics provides a comprehensive view of a model's ability to retain, transfer, and adapt knowledge in a sequential learning setting:

- **Average Incremental Score** (AIS) is a straightforward metric that evaluates the model's overall performance:

$$AIS = \frac{1}{T}\sum_{i=1}^{T}\frac{1}{i}\sum_{j=1}^{j} ia_{i,j}$$

As we use F1-score as initial score, the AIS ranges from 0 (worst performance) to 100 (best performance).

- **Backward Transfer** (BWT) measures how learning new domains influences performance on previously learned domains. A positive BWT indicates that new learning improves past knowledge, while a negative BWT

CLP	Hyperparameter	Grid Search	Learning rate	Batch size	Epochs
EWC	$\lambda = 100$	[1, 5, 10, 100, 200, 500]			
SI	$\lambda = 0.5$	[0.1, 0.5, 1, 10]			
MAS	$\lambda = 200$	[10, 100, 200, 300]	$2e^{-5}$	64	3
A-GEM	$M = 200$	None			
NER	$M = 200$	None			

Table 3: Model hyper-parameters for RoBERTa$_{3Tasks}$ and FlauBERT$_{FineTuned_3Tasks}$.

indicates catastrophic forgetting:

$$BWT = \frac{1}{T-1} \sum_{i=1}^{T-1} (a_{T,i} - a_{i,j})$$

The objective is to maximize BWT.

- **Forgetting Measure (FM)** pictures how much the model forgets its maximum performance seen on each domain. A positive FM means catastrophic forgetting while negative FM indicates strong knowledge transfer:

$$FM = \frac{1}{T-1} \sum_{i=1}^{T-1} (max_{1 \leq j \leq T-1}(a_{j,i}) - a_{T,i})$$

In this case, the objective is to minimize FM.

- **LAST** score indicates the model's last performances:
$LAST = \frac{1}{T}\ sum_{i=1}^{T-1} a_{T,i}$

As we use F1-score as initial score, the LAST score ranges from 0 (worst performance) to 100 (best performance).

Among these measures, two performance metrics, AIS and LAST, show respectively how the model performs during all the process and how a model performs once it reaches its final state. Both transfer knowledge indicators, BWT and FM, represent respectively how much the model is efficient compared to previous steps and how many points of F1-score are lost against the best step among all previous steps.

5. Results

5.1. Continual Learning vs. Baselines

Tables 4 and 5 present the results on the French and English datasets respectively. The best results for each type of method are in bold and the best result for each metric is underlined.

For Kozlowski dataset, the upper limit of the *cumulative* baseline remains unrivaled on the three classification tasks, reaching the best LAST, the only negative FM (i.e. virtually no forgetting) and the highest positive BWT. Thus, this baseline model is the most efficient model at the last step, learns the most from previous steps and forgets less. However, regularization methods (EWC, SI, MAS) improve upon vanilla fine-tuning in FM (2.35, 2.48, 2.94 vs. 3.36) which means that continual learning methods forget less information than no continual learning protocol. MAS even outperfoms cumulative on AIS, meaning that even if at the end a cumulative method outperforms all continual learning protocols, at early stages, continual learning, especially MAS, manages to get the best results. Replay-based baselines (NER, A-GEM) perform less on LAST (47.94 for best replay method and 48.28 for best regularization method), nor reduce forgetting substantially (FM 2.19 vs 2.35 for regularization, which is only a gain of 0.26).

On HumAID dataset, the best regularization method (EWC) surpasses the cumulative method on almost every axis. EWC attains the highest LAST for all tasks (64.23, 73.92, 75.70) and displays the most negative FM (-2.50, -1.43, -0.44), signaling effective forgetting mitigation. MAS shows a similar trend, with the strongest positive BWT on humanitarian and utility tasks. The humanitarian task being more complex for English (9 labels vs. 7 labels for French) with a highly imbalanced dataset, a continual learning protocol will learn from the most frequent labels (e.g for flood, HELP, OTHER and SUPPORT) and will freeze these important weights to not forget how to handle these labels while learning the representative labels of new crisis types (e.g. HUMAN DAMAGE for fire). Once again, replay methods perform less than regularization approaches. Indeed according to recent CL surveys (Wang et al., 2024), replay baselines are better (and consume more resources) with a higher number of domains while regularization benefits from close domains as they can easily protect the most important parameters. Our findings corroborate these ideas: when domains are small and topically related, regularization-based methods can exploit shared knowledge from different domains while protecting critical weights, outperforming naive fine-tuning.

5.2. Impact of Crisis Types

As continual learning were less efficient on Kozlowski dataset, we conducted an additional experiment with EWC in order to identify which type

Type	CLP	Humanitarian				Urgency				Utility			
		AIS	LAST	FM	BWT	AIS	LAST	FM	BWT	AIS	LAST	FM	BWT
Baselines	Cumulative	46.69	**50.36**	**-1.28**	**7.39**	**64.39**	**66.48**	**-0.27**	**4.27**	77.61	**78.75**	**0.78**	**2.04**
	Vanilla	**47.12**	47.97	3.36	2.11	63.78	64.21	2.80	0.41	**77.76**	77.85	2.51	0.17
Regularization	EWC	46.38	47.13	2.94	**2.59**	**64.56**	**64.95**	**2.86**	**1.02**	78.04	**78.51**	2.03	**0.51**
	SI	47.08	48.12	**2.35**	2.34	64.27	64.87	2.97	0.64	77.77	77.95	2.53	0.17
	MAS	**47.25**	**48.28**	2.48	2.34	63.71	64.32	3.24	0.74	**78.16**	78.50	**1.75**	0.42
Replay	NER	**46.93**	**47.94**	**2.19**	**3.77**	**64.23**	**64.74**	**2.91**	**1.74**	**77.62**	**77.81**	**1.86**	0.04
	A-GEM	44.77	45.61	3.23	3.10	62.05	62.40	3.25	1.21	75.92	76.14	2.81	**0.46**

Table 4: Results for Kozlowski dataset with FlauBERT$_{FineTuned_3Tasks}$ on humanitarian, urgency and utility tasks.

Type	CLP	Humanitarian				Urgency				Utility			
		AIS	LAST	FM	BWT	AIS	LAST	FM	BWT	AIS	LAST	FM	BWT
Baselines	Cumulative	**61.67**	**64.10**	-1.31	7.83	**71.26**	**73.02**	**-0.37**	**4.79**	**73.90**	**74.67**	**0.56**	2.00
	Vanilla	59.40	62.20	**-1.47**	**7.93**	70.37	71.92	-0.20	4.77	73.41	73.69	1.44	**2.02**
Regularization	EWC	**60.63**	**64.23**	-2.50	9.21	**71.76**	**73.92**	**-1.43**	5.67	**74.71**	**75.70**	**-0.44**	3.19
	SI	60.35	63.44	-1.55	8.55	71.23	72.93	-0.16	4.84	73.89	74.22	1.14	2.67
	MAS	60.13	64.02	**-3.02**	**9.53**	71.35	73.57	-1.27	**6.04**	74.53	75.63	-0.27	**3.37**
Replay	NER	**61.91**	**64.09**	**-0.97**	**7.34**	**71.57**	**72.95**	**-0.33**	4.26	**74.39**	**74.69**	0.64	1.91
	A-GEM	61.44	63.44	-0.58	7.13	71.03	72.49	-0.07	**5.32**	74.27	74.55	**0.58**	**2.55**

Table 5: Results for HumAID dataset wit RoBERTa_$3Tasks$ on humanitarian, urgency and utility tasks.

of crisis has an impact on the performance. For this purpose, we rely on previous work from Meunier et al. (2025) showing that *sudden* crises (crises which can hardly be predicted, e.g. earthquake, fire) are more difficult to deal with than *expected* crises (crises which can be meteorologically predicted, e.g. flood, hurricane).

Thus, we designed a CL protocol, called SUDDEN-EXPECTED, with EWC for 2 experiments: training on 2 *expected* (resp. *sudden*) crisis types and testing on 1 *sudden* (resp. *expected*) crisis type, in order to have *sudden* and *expected* data with comparable size (see Table 1). We used AIS as an evaluation score. Then, we evaluate the impact of crisis types on the performance by computing for each tested crisis type the difference between the AIS score obtained via the CrisisCL protocol and the AIS score obtained via the SUDDEN-EXPECTED protocol. Indeed, a low or negative difference means that the SUDDEN-EXPECTED protocol has better results and that the model heavily relies on the crisis types used for training. On the contrary, a positive difference means that the CrisisCL protocol has better results and that the crisis types used for training do not bring enough information to deal with a new type of crisis. The results are presented in Table 6.

We observe that for all expected crises, there is a huge drop of performance on the 3 tasks when using the SUDDEN-EXPECTED protocol (positive difference). Concerning sudden crises, the drop is low for *Fire*, due to a fuzzy crisis type covering different event types with different characteristics such as wildfire, structure fire, etc. Moreover, although crisis types such as flood, hurricanes and storm share specific characteristics (water related damages, predictability, large scale crises) and similar vocabulary, the vocabulary is different for *Fire*.

Crisis	Humanitarian	Urgency	Utility
EXPECTED			
Hurricane	22.0	21.1	12.6
Storms	26.5	23.5	20
Flood	23.8	20.7	7.6
SUDDEN			
Fire	2.9	2.85	3.79
Collapse	0.01	-2.55	-5.31

Table 6: Difference between AIS obtained by CrisisCL and Sudden-Expected with EWC on Kozlowski dataset.

When the model is trained on these crisis types, while continual learning method try to prevent forgetting of fire crisis, the amount of data tend to prioritize flood, storm and hurricane. The only crisis which benefits from the SUDDEN-EXPECTED protocol is *Collapse*. This is due to a shared vocabulary with expected crises (e.g. the following tweet about Irma hurricane relating a building collapse: *#Irma has caused #BVI infrastructure to collapse more support is needed from major aid orgs & government*). These results confirm what has been shown in previous work, i.e. a good quality of expected crises data with valuable information that a model can learn while sudden crises are more difficult to handle. These results emphasis the need for online training. Indeed, without a good representation of all crisis types, there is a possibility that the model lacks the information required to perform on a new domain. Moreover, all crisis data do not have an equivalent quality (number of tweets, distribution, variety of textual data) and with a step by step training, we are able to identify "noisy" crises.

6. Conclusion

In this paper, we presented CrisisCL, a benchmark for continual learning in crisis management which provides a structured protocol for Domain-Incremental Learning in this context, highlighting the trade-offs between memory-based and regularization-based approaches under strict resource constraints. The results confirm that while regularization methods such as EWC and MAS can effectively mitigate catastrophic forgetting — especially on large, topically coherent datasets —, replay-based methods often struggle when memory budgets are limited. Furthermore, the additional protocols (e.g., Out-of-Type and Sudden-Expected) revealed the difficulty to generalize to unseen crisis types, emphasizing the need for robust domain adaptation and domain-aware architectures.

Future work will explore the integration of adaptive weighting of domains, and the implementation of new continual learning scenarios such as Class Incremental Learning for crisis management.

Acknowledgments

We would like to thank Angela Yao and Guodong Ding for their help on continual learning strategies. This work has been supported by DesCartes: the National Research Foundation, Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) program.

Limitations

While CrisisCL is an efficient benchmark for domain incremental learning in crisis management, it can be expanded to cover more scenarios such as task incremental learning or instance incremental learning. The experimentation can also be extended on more type of crises or other languages.

Ethics Statement

The dataset used in this paper are textual datasets publicly available to the research community. The datasets are anonymized and contain no offensive or abusive language. They were collected before Twitter changed to X and conform to the Twitter Developer Agreement and Policy that allows unlimited distribution of either the numeric identification number or the textual content of each tweet.

7. Bibliographical References

Manoj Acharya, Tyler L Hayes, and Christopher Kanan. 2020. Rodeo: Replay for online object detection. *arXiv preprint arXiv:2008.06439*.

Firoj Alam, Ferda Ofli, and Muhammad Imran. 2018. CrisisMMD: Multimodal Twitter Datasets from Natural Disasters. In *Proceedings of the 12th International AAAI Conference on Web and Social Media (ICWSM)*.

Nilani Algiriyage, Rangana Sampath, Raj Prasanna, Emma EH Doyle, Kristin Stock, and David Johnston. 2021. Identifying disaster-related tweets: a large-scale detection model comparison. In *Social Media in Crises and Conflicts, Proceedings of the 18th ISCRAM Conference*, pages 731–743.

Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. 2018. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European conference on computer vision (ECCV)*, pages 139–154.

Rahaf Aljundi, Eugene Belilovsky, Tinne Tuytelaars, Laurent Charlin, Massimo Caccia, Min Lin, and Lucas Page-Caccia. 2019a. Online continual learning with maximal interfered retrieval. *Advances in neural information processing systems*, 32.

Rahaf Aljundi, Klaas Kelchtermans, and Tinne Tuytelaars. 2019b. Task-free continual learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11254–11263.

Rahaf Aljundi, Min Lin, Baptiste Goujaud, and Yoshua Bengio. 2019c. Gradient based sample selection for online continual learning. *Advances in neural information processing systems*, 32.

Jihwan Bang, Hyunseo Koh, Seulki Park, Hwanjun Song, Jung-Woo Ha, and Jonghyun Choi. 2022. Online continual learning on a contaminated data stream with blurry task boundaries. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9275–9284.

Alexandre Berard. 2021. Continual learning in multilingual NMT via language-specific embeddings. In *Proceedings of the Sixth Conference on Machine Translation*, pages 542–565, Online. Association for Computational Linguistics.

Jürgen Beringer and Eyke Hüllermeier. 2007. Efficient instance-based learning on data streams. *Intelligent Data Analysis*, 11(6):627–650.

Nils Bourgon, Farah Benamara, Alda Mari, Véronique Moriceau, Gaetan Chevalier, and Laurent Leygue. 2022. Are Sudden Crises Making me Collapse? Measuring Transfer Learning Performances on Urgency Detection. In *19th International Conference on Information Systems for Crisis Response and Management (ISCRAM 2022)*.

Cornelia Caragea, Adrian Silvescu, and Andrea Tapia. 2016. Identifying Informative Messages in Disasters Events using Convolutional Neural Networks. In *13th International Conference on Information Systems for Crisis Response and Management (ISCRAM 2016)*, pages 1–7.

Carlos Castillo. 2016. *Big crisis data: social media in disasters and time-critical situations*. Cambridge University Press.

Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet K Dokania, Philip HS Torr, and Marc'Aurelio Ranzato. 2019. On tiny episodic memories in continual learning. *arXiv preprint arXiv:1902.10486*.

Jishnu Ray Chowdhury, Cornelia Caragea, and Doina Caragea. 2020. On Identifying Hashtags in Disaster Twitter Data. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI)*.

Yung-Sung Chuang, Shang-Yu Su, and Yun-Nung Chen. 2020. Lifelong language knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2914–2924, Online. Association for Computational Linguistics.

Jens A. de Bruijn, Hans de Moel, Albrecht H. Weerts, Marleen C. de Ruiter, Erkan Basar, Dirk Eilander, and Jeroen C.J.H. Aerts. 2020. Improving the classification of flood tweets with contextual hydrological information in a multimodal neural network. *Computers & Geosciences*, 140:104485.

Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. 2021. A continual learning survey: Defying forgetting in classification tasks. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3366–3385.

Alexis Dusart, Karen Pinel-Sauvagnat, and Gilles Hubert. 2021. ISSumSet: A Tweet Summarization Dataset Hidden in a TREC Track. In *Proceedings of the 36th Annual ACM Symposium on Applied Computing*, SAC '21, page 665–671, New York, NY, USA. Association for Computing Machinery.

Mansi Jain, Harmeet Kaur, Bhavna Gupta, Jaya Gera, and Vandana Kalra. 2025. Incremental learning algorithm for dynamic evolution of domain specific vocabulary with its stability and plasticity analysis. *Scientific Reports*, 15(1):272.

Marc-André Kaufhold, Markus Bayer, and Christian Reuter. 2020. Rapid relevance classification of social media posts in disasters and emergencies: A system and evaluation featuring active, incremental and online learning. *Information Processing & Management*, 57(1):102132.

Zixuan Ke and Bing Liu. 2022. Continual learning of natural language processing tasks: A survey. *arXiv preprint arXiv:2211.12701*.

Ronald Kemker and Christopher Kanan. 2017. Fearnet: Brain-inspired model for incremental learning. *ArXiv*, abs/1711.10563.

Jens Kersten, Anna Kruspe, Matti Wiegmann, and Friederike Klan. 2019. Robust Filtering of Crisis-related Tweets. In *ISCRAM 2019 conference proceedings-16th international conference on information systems for crisis response and management*.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526.

Rani Koshy and Sivasankar Elango. 2023. Multimodal tweet classification in disaster response systems using transformer-based bidirectional attention model. *Neural Computing and Applications*, 35(2):1607–1627.

Jędrzej Kozal, Jan Wasilewski, Bartosz Krawczyk, and Michał Woźniak. 2024. Continual learning with weight interpolation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4187–4195.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier, and Didier Schwab. 2020. FlauBERT: Unsupervised language model pre-training for French. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France. European Language Resources Association.

Hongmin Li, Doina Caragea, Cornelia Caragea, and Nic Herndon. 2018. Disaster response aided

by tweet classification with a domain adaptation approach. *Journal of Contingencies and Crisis Management*, 26(1):16–27.

Zhizhong Li and Derek Hoiem. 2017. Learning without forgetting.

Junhua Liu, Trisha Singhal, Lucienne T.M. Blessing, Kristin L. Wood, and Kwan Hui Lim. 2021. CrisisBERT: A Robust Transformer for Crisis Classification and Contextual Crisis Embedding. In *Proceedings of the 32nd ACM Conference on Hypertext and Social Media*, HT '21, page 133–141, New York, NY, USA. Association for Computing Machinery.

David Lopez-Paz and Marc'Aurelio Ranzato. 2022. Gradient episodic memory for continual learning.

Sreenivasulu Madichetty, Sridevi Muthukumarasamy, and Sreekanth Madisetty. 2023. A RoBERTa based model for identifying the multi-modal informative tweets during disaster. *Multimedia Tools and Applications*.

Zheda Mai, Ruiwen Li, Jihwan Jeong, David Quispe, Hyunwoo Kim, and Scott Sanner. 2022. Online continual learning in image classification: An empirical survey. *Neurocomputing*, 469:28–51.

Richard McCreadie and Cody Buntain. 2023. Crisisfacts: building and evaluating crisis timelines. *International ISCRAM Conference*.

Richard McCreadie, Cody L. Buntain, and Ian Soboroff. 2019. Trec incident streams: Finding actionable information on social media. In *International Conference on Information Systems for Crisis Response and Management*.

Romain Meunier, Farah Benamara, Véronique Moriceau, Zhongzheng Qiao, and Savitha Ramasamy. 2025. CrisisTS: Coupling Social Media Textual Data and Meteorological Time Series for Urgency Classification. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16082–16099, Vienna, France. Association for Computational Linguistics.

Romain Meunier, Farah Benamara, Véronique Moriceau, and Patricia Stolf. 2023. Image and Text: Fighting the Same Battle? Super-resolution Learning for Imbalanced Text Classification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Findings of the Association for Computational Linguistics: EMNLP 2023, pages 10707–10720, Singapour, Singapore. Association for Computational Linguistics.

Umberto Michieli and Mete Ozay. 2024. Hop to the next tasks and domains for continual learning in nlp. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 14359–14369.

Venkata Kishore Neppalli, Cornelia Caragea, and Doina Caragea. 2018. Deep Neural Networks versus Naive Bayes Classifiers for Identifying Informative Tweets during Disasters. In *Proceedings of the 15th International Conference on Information Systems for Crisis Response and Management*, ISCRAM'2018.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly

Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report.

Hakan T Otal and M Abdullah Canbaz. 2024. LLM-Assisted Crisis Management: Building Advanced LLM Platforms for Effective Emergency Response and Public Collaboration. *arXiv preprint arXiv:2402.10908*.

Aman Priyanshu, Mudit Sinha, and Shreyans Mehta. 2021. Continual distributed learning for crisis management. *arXiv preprint arXiv:2104.12876*.

Faghihi Hossein Rajaby, Bashar Alhafni, Ke Zhang, Shihao Ran, Joel Tetreault, and Alejandro Jaimes. 2022. CrisisLTLSum: A benchmark for local crisis event timeline extraction and summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5455–5477, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. 2017. icarl: Incremental classifier and representation learning.

Christian Reuter, Amanda Lee Hughes, and Marc-André Kaufhold. 2018. Social Media in Crisis Management: An Evaluation and Analysis of Crisis Informatics Research. *International Journal of Human–Computer Interaction*, 34(4):280–294.

Roberta Rocca, Nicolò Tamagnone, Selim Fekih, Ximena Contla, and Navid Rekabsaz. 2023. Natural language processing for humanitarian action: Opportunities, challenges, and the path toward humanitarian NLP. *Frontiers in Big Data*, 6.

Shrey Satapara and P. K. Srijith. 2024. TL-CL: Task and language incremental continual learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12123–12142, Miami, Florida, USA. Association for Computational Linguistics.

Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. 2017. Continual learning with deep generative replay. In *Neural Information Processing Systems*.

Fan-Keng Sun, Cheng-Hao Ho, and Hung-Yi Lee. 2019. Lamol: Language modeling for lifelong language learning.

Reem Suwaileh, Tamer Elsayed, and Muhammad Imran. 2023. IDRISI-RE: A generalizable dataset with benchmarks for location mention recognition on disaster tweets. *Information Processing & Management*, 60(3):103340.

Cagri Toraman, Izzet Emre Kucukkaya, Oguzhan Ozcelik, and Umitcan Sahin. 2023. Tweets Under the Rubble: Detection of Messages Calling for Help in Earthquake Disaster. *arXiv preprint arXiv:2302.13403*.

Gido M Van de Ven and Andreas S Tolias. 2019. Three scenarios for continual learning. *arXiv preprint arXiv:1904.07734*.

Congcong Wang, Paul Nulty, and David Lillis. 2021. Transformer-based Multi-task Learning for Disaster Tweet Categorisation. In *International Conference on Information Systems for Crisis Response and Management*.

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. 2024. A comprehensive survey of continual learning: Theory, method and application. *IEEE transactions on pattern analysis and machine intelligence*, 46(8):5362–5383.

Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, et al. 2022. Dualprompt: Complementary prompting for rehearsal-free continual learning. In *European conference on computer vision*, pages 631–648. Springer.

Friedemann Zenke, Ben Poole, and Surya Ganguli. 2017. Continual learning through synaptic intelligence. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3987–3995. PMLR.

Peng Zhang, Byron J. Gao, Xingquan Zhu, and Li Guo. 2011. Enabling fast lazy learning for data streams. In *2011 IEEE 11th International Conference on Data Mining*, pages 932–941.

Zhi-Hua Zhou. 2022. Open-environment machine learning. *National Science Review*, 9(8):nwac123.

8. Language Resource References

Firoj Alam, Umair Qazi, Muhammad Imran, and Ferda Ofli. 2021. HumAID: human-annotated disaster incidents data from Twitter with deep learning benchmarks. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 933–942.

Diego Kozlowski, Elisa Lannelongue, Frédéric Saudemont, Farah Benamara, Alda Mari, Véronique Moriceau, and Abdelmoumene Boumadane. 2020. A three-level classification of French tweets in ecological crises. *Information Processing & Management*, 57(5).