Image and Text: Fighting the Same Battle?
Super-resolution Learning for Imbalanced Text Classification

Romain Meunier[1] and **Farah Benamara**[1,2] and **Véronique Moriceau**[1] and **Patricia Stolf**[1]
(1) IRIT, Université de Toulouse, *CNRS, Toulouse INP, UT3*, Toulouse, France
(2) IPAL, CNRS-NUS-ASTAR, Singapore
firstname.lastname@irit.fr

Abstract

In this paper, we propose SRL4NLP, a new approach for data augmentation by drawing an analogy between image and text processing: Super-resolution learning. This method is based on using high-resolution images to overcome the problem of low resolution images. While this technique is a common usage in image processing when images have a low resolution or are too noisy, it has never been used in NLP. We therefore propose the first adaptation of this method for text classification and evaluate its effectiveness on urgency detection from tweets posted in crisis situations, a very challenging task where messages are scarce and highly imbalanced. We show that this strategy is efficient when compared to competitive state-of-the-art data augmentation techniques on several benchmarks datasets in two languages.

1 Introduction

The performances of NLP classification tasks are largely influenced by the availability of good quality training data with sufficient number of representative instances. This is particularity salient for (highly) imbalanced text classification problems, like malware detection (Oak et al., 2019), relation extraction (Gao et al., 2019), hate speech detection (Schmidt and Wiegand, 2017) as well as urgency detection (Sánchez et al., 2023).

To augment minority classes, many data augmentation strategies have been proposed while ensuring models generalization. Roughly, these strategies can operate either at the data or model level. Data-based approaches generate synthetic new instances relying on lexical/syntactic variations (paraphrasing, substitution, etc.) or pre-trained generative models (e.g., GPT2, GPT3) (Anaby-Tavor et al., 2020; Yoo et al., 2021; Chia et al., 2022). Model-based approaches on the other hand fight class imbalance during training via (a) some dedicated loss functions by assigning higher wrong classification costs to classes with small proportion (Lin et al., 2020; Li et al., 2020), (b) specific architectures such as multi-task learning leveraging auxiliary tasks (Spangher et al., 2021) and knowledge distillation that trains smaller model with soft labels (Tan et al., 2022), or (c) text interpolation that mixes up original and perturbed samples in the hidden space (Chen et al., 2022). While data-based techniques are easy to implement with low computational cost compared to model-based ones (Wei and Zou, 2019), current methods have difficulty to generate meaningful instances as augmented data are often too close to the original ones, or semantically different which may result in instances that do not make sense to humans (Feng et al., 2021; Kamalloo et al., 2022). Prompt-based augmentation alleviates these issues but at the price of training large-scale language models. In addition, they turn to be sensitive to prompt design in terms of the choice and order of the input examples (Reynolds and McDonell, 2021).

In this paper, we propose a new approach for data augmentation by drawing an analogy between image and text processing: *Super-resolution learning* (Yang et al., 2019). This method is based on using high-resolution images to overcome the problem of low resolution images and has been successfully employed in image processing when images have a low resolution or are too noisy (Peng et al., 2016; Koziarski and Cyganek, 2018; Tian et al., 2023; Fukami et al., 2023). It has however never been adapted to NLP classification tasks. In this context, we make a parallel between image and text by considering noisy and short user generated content (e.g., social media posts) as a low resolution image and long well written documents (e.g., news papers, Wikipedia, scientific articles) as a high resolution image which enables augmentation with more diverse and high quality data with less cost. Mixed data from both types of resolution are then used to train supervised classifiers.

We measure the effectiveness of super-resolution data augmentation on single label text classification tasks which: (i) show a step-imbalanced distribution with a high fraction of minority classes (Buda et al., 2018), (ii) are characterized by a long-tailed distribution with many instances for a small number of classes, but few for the remaining classes (Henning et al., 2023), (iii) involve low-quality instances (e.g., few tokens, ill-formed with grammar spelling/syntactic issues), and (iv) are about media events such as disasters (e.g., flood, earthquakes), health (e.g., vaccines, cancer, suicide) and sport, that are more likely to be discussed in factual external sources of information, adding therefore meaningful context. Among NLP applications meeting these criteria (e.g., cybersecurity (Oak et al., 2019), medecine (Mayer et al., 2021; Ansari et al., 2021)), we experiment with urgency detection from tweets (Castillo, 2016; Alam et al., 2021), a very challenging task aiming at classifying messages posted during crisis situations according to fine-grained intention to act categories (such as human/infrastructure damages, warning/advice). In this task, messages that require an urgent and rapid action from rescue teams are scarce and highly imbalanced compared to non-urgent ones (Reuter et al., 2018; Sarioglu Kayi et al., 2020), making it a very interesting case study for super-resolution learning. In this paper, our contributions are:

- SRL4NLP, a simple yet effective adaptation of a well-established data augmentation strategy in image processing to text classification which does not rely on artificially generated data but on existing high quality texts.

- A study of the performances of this strategy on disaster management from social media relying on benchmark datasets in both English and French.

- A quantitative and qualitative evaluation in an out-of-type configuration where classifiers are evaluated in unseen event types during training. Our results suggest that super-resolution learning is achieving competitive results when compared to state-of-the-art data augmentation techniques.

In the following, Section 2 presents the original machine-learned super-resolution method and how we adapted it for NLP. Section 3 presents the datasets and Section 4 presents the experimental settings used in this study. Section 5 presents the results. We compare our approach to related work in Section 6. Finally, we conclude drawing some perspectives for future work.

2 Data Augmentation with Super-Resolution

2.1 Super-resolution for Image Processing

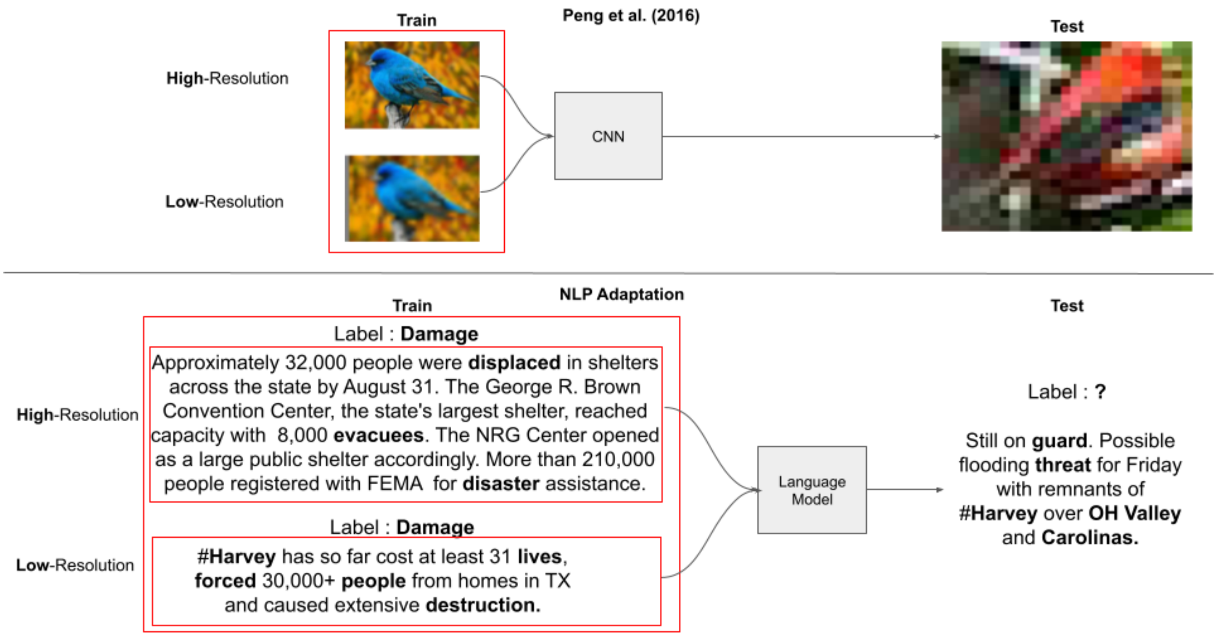
Data augmentation has first been explored in the field of image processing with algorithms for geometric transformations, color changes, filters, random erasing, etc. (see (Shorten and Khoshgoftaar, 2019) for a survey). Classifiers are mainly trained on high-resolution labeled data but do not perform well when tested on low-resolution data. Thus, super-resolution method, consisting in enhancing the resolution of an image, is largely approved and applies in many fields such as face recognition (Zou and Yuen, 2012), microscopic image (Nehme et al., 2018), the domains of flow (Liu et al., 2020; Yousif et al., 2021; Fukami et al., 2023) or medicine (van Sloun et al., 2019).

Super-resolution gained interest for deep learning with Peng et al. (2016) who evaluated different training strategies for CNN models to improve fine-grained category classification of very low resolution images. They artificially reduce the resolution of some high-resolution training data and show that mixing them during the training phase improves accuracy when models are tested on low-resolution images compared to training on either high or low-resolution only.

Thus, multi-resolution training, which randomly resizes images to different resolutions, allows to accommodate varying resolutions during testing and is used mainly to train CNN models. Recently, Tian et al. (2023) proposed ResFormer, a vision transformer trained on multi-resolution images (several scalings of images) which obtains better results than a transformer trained on single resolution.

2.2 SRL4NLP: Super-resolution for NLP

Inspired by these results, we draw an analogy between images and texts: (1) we consider that tweets are low-resolution texts as they are short and often low-quality, noisy texts; and (2) we propose to augment low-resolution training data with similar texts with a higher resolution (see Figure 1). We consider high-resolution texts as clean texts with a higher number of tokens than tweets. Instead of generating artificial high-resolution texts from low-



Approximately 32,000 people were **displaced** in shelters across the state by August 31. The George R. Brown Convention Center, the state's largest shelter, reached capacity with 8,000 **evacuees**. The NRG Center opened as a large public shelter accordingly. More than 210,000 people registered with FEMA for **disaster** assistance.

#Harvey has so far cost at least 31 **lives**, **forced** 30,000+ **people** from homes in TX and caused extensive **destruction.**

Still on **guard**. Possible flooding **threat** for Friday with remnants of **#Harvey** over **OH Valley** and **Carolinas.**

Figure 1: Adaptation of Peng et al. (2016) super-resolution learning to text classification, illustrated here on the task of urgency detection, our case study.

resolution ones before or during training as often done in image processing and to avoid additional training costs, we propose to augment data with existing texts such as newspaper, encyclopedic articles, or scientific publications, and to mix them with tweets for the training step.

SRL4NLP implies to have a high-resolution dataset about the same topics as the low-resolution one, and annotated with the same labels. If not available, we propose a method in Section 3 to collect and automatically annotate high-resolution data from structured texts, such as Wikipedia articles.

3 Datasets

The literature on emergency detection has been growing fast in the recent years and several datasets (mainly tweets) have been proposed to account for crisis-related phenomena.[1] Messages are annotated according to relevant categories that are deemed to fit the information needs of various stakeholders like humanitarian organizations, local police and firefighters. Well-known dimensions include relatedness to identify whether the message content is useful (Jensen, 2012) and situation awareness (also known as urgency) to filter out on-topic relevant vs. on-topic irrelevant information (Imran et al., 2013; McCreadie et al., 2019; Sarioglu Kayi et al., 2020; Kozlowski et al., 2020). We evaluate several data augmentation strategies on different datasets in English and French since datasets for crisis management are available in these languages (see Table 1).

3.1 Low-Resolution datasets

- KOZLOWSKI$_{merge}$: This dataset is the largest French dataset for crisis management composed of 16,202 tweets about various expected (e.g., storms) and sudden crisis (e.g., explosions) events (Kozlowski et al., 2020; Bourgon et al., 2022). It contains tweets annotated for three urgency categories as well as 6 intentions to act 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 on actionability but contribute in raising situational awareness, and finally (3) NOT USEFUL for messages that are not related to the targeted crisis. To be comparable with the following English dataset, some intent classes have been merged into more coarse-grained classes: HUMAN-DAMAGE and MATERIAL-DAMAGE into DAMAGE; SUPPORT and CRITICS into SOCIAL while

[1] See https://crisisnlp.qcri.org/ for an overview.

	Urgent		Not Urgent		Not Useful	Total
	Damage	Advice_Warning	Social	Other		
Kozlowski$_{merge}$	908	1,456	1,128	1,352	11,358	**16,202**
HumAID$_{merge}$	908	1,456	1,128	1,352	11,358	**16,202**
FR$_{AUTO}$	603	276	153	0	0	**1,032**
FR$_{AUTO\_NO\_SOCIAL}$	603	276	0	0	0	**879**
EN$_{AUTO\_NO\_SOCIAL}$	1,160	0	0	0	0	**1,160**

Table 1: Distribution of each dataset in terms of urgency (3 classes) and intent (5 classes).

keeping imbalanced distribution with 71%, 8,98% and 5.60% for Not Useful, Social and Damage respectively.

- **HumAID**$_{merge}$: It is an English dataset of 77,196 tweets annotated for crisis management following 10 intent categories and one Not Useful class (Alam et al., 2021). This latter class having few instances (1,986) and as we wanted it to be a majority class as in Kozlowski$_{merge}$, we added 9,372 Not Useful instances from other crisis datasets annotated for relatedness (CrisisLexT6 (Olteanu et al., 2014) and CrisisLexT26 (Olteanu et al., 2015)). To allow a better comparison, we aligned HumAID annotations with Kozlowski$_{merge}$ by merging similar categories (e.g., injured/dead people and missing or found people into the unified label Damage) while discarding irrelevant ones (e.g., no decision). We have also randomly selected the same number of tweets from the same classes as in Kozlowski$_{merge}$ to have an equivalent distribution and a comparable size.

3.2 High-Resolution datasets

SRL4NLP implies to have a high-resolution dataset about the same topics as the low-resolution one, and annotated with the same labels, so that both datasets can be mixed during the training step. However, as there is no available high-resolution dataset for crisis management which fit our needs, we defined a simple but generic method to collect and annotate automatically high-resolution texts from existing structured documents. This method avoids additional computational costs due to artificial data generation and manual annotation. The selection and annotation of high-resolution texts from structured documents follow Algorithm 1, which annotates paragraphs of relevant documents according to the title of the section they belong to; assuming a set of labels C of a classification task for a given use case and a set of documents D related to this use case structured in terms of sections S and paragraphs P, such that:

- $C = \{(c_1, k_1), ..., (c_n, k_n)\}$ where each label c_i is triggered by a set of keywords $K = \{k_1, ..., k_l\}$, and
- $D = \{d_1, ..., d_m\}$ where $d_j =< (S_j, P_i^+)^+ >$ such that $S_j = \{s_{j,1}, ..., s_{j,n}\}$ the n sections of $d_j \in D$, and $P_n = \{p_{n,1}, ..., p_{n,r}\}$ the r paragraphs of S_j.

Algorithm 1: High-resolution text scrapper and annotation

```
Input: $C$, $K$, $D$, $S$ , $P$
Output: label($p_r$)
for $d_i$ in $D$ do
    for $s_{i,j}$ in $S_i$ do
        for $c_n$,$k_n$ in $C$ do
            if $k_n$ in $s_j$ then
                for $p_r$ in $P_j$ do
                    label($p_r$) = $c_n$
```

For our use case, we decided to use Wikipedia as external source as it contains a large variety of factual and valuable information for understanding the nature, causes, impacts, and responses to different crises (e.g., damage and weather reports, responses and mitigation efforts, media coverage and public reactions) making it well suited for super-resolution. In addition, it is copyright free, easy to scrap, and comes with a structured format with dedicated sections and paragraphs for each Wikipedia page. Thus, the following datasets, built with Algorithm 1, are composed of high-resolution paragraphs from Wikipedia, used to augment all minority classes, except Not Useful, the majority class:

- **FR**$_{AUTO}$: the set of relevant documents D is composed of the 152 links in the French Wikipedia page *Natural disasters by year*[2]. The sets of keywords K used for the matching of section titles

[2] https://fr.wikipedia.org/wiki/Cat%C3%A9gorie:Catastrophe_naturelle_par_ann%C3%A9e

with classes are given in Appendix A. This results in 1,032 labelled paragraphs, each of them being related to a crisis type given by the document title. Among them, we manually checked 144 randomly selected paragraphs and agreed with the automatic annotation (see an example in Appendix C).

- **FR**$_{AUTO\_NO\_SOCIAL}$: unlike Twitter, texts in Wikipedia which can be labelled as SOCIAL (i.e. critics, support) are rare. For this reason, we have created a subset of FR$_{AUTO}$, named FR$_{AUTO\_NO\_SOCIAL}$, which contains only the data labelled DAMAGE and ADVICE_WARNING.

- **EN**$_{AUTO\_NO\_SOCIAL}$: in the same way, 213 links in the English Wikipedia page *Natural disasters by year*[3] have been scrapped leading to the automatic selection and annotation of 1,160 paragraphs as DAMAGE or ADVICE_WARNING (see Appendix B for the English matching lexicon).

Table 1 shows the distribution of these augmented datasets.

4 Models and Evaluation Protocol

4.1 Experimental Settings

For testing the datasets in French, we used **FlauBERT**$_{Fine-Tuned}$ which is the best performing model for intent classification on the French dataset according to (Kozlowski et al., 2020). It is a FlauBERT model (Le et al., 2020) that has been fine-tuned on 358,834 unlabelled tweets posted during crises. For English, we used **RoBERTa** (Liu et al., 2019) which is reported to be efficient for crisis management in English (Koshy and Elango, 2022; Rocca et al., 2023; Madichetty et al., 2023). For both models, we used the Focal Loss with an Adam optimizer.

Regarding the classification task, we do not report any comparative analysis with existing mainstream large language model such as GPT for three main reasons. First, our aim is to design a simple approach that requires very low computational cost. Second, as we used GPT3 to generate the new augmented data and so learned from these data, we decided to not use it for classification because results may be biased. Finally, we also experimented classification with GPT3 on 100 tweets of KOZLOWSKI$_{merge}$. However, it was unable to distinguish OTHER and NOT USEFUL messages, achieving therefore low performances compared to FlauBERT (e.g. GPT3 was only able to correctly classify 35% of messages into intention to act categories).

Following the general trends in crisis management (Kersten et al., 2019; Algiriyage et al., 2021; Bourgon et al., 2022), we designed an *out-of-type* evaluation protocol by training on a pool of events related to different types of crises (e.g., Hurricane, Storm) and testing on a particular different type (e.g., Earthquake). The aim is to evaluate if a model can deal with new types of crisis, which is crucial to ensure the portability of the models to unseen events.

Each corpus contains different types of crisis: Fire, Flood, Storm, Hurricane, Collapse and Terrorist attack for KOZLOWSKI$_{merge}$ and Fire, Flood, Storm, Hurricane, Earthquake for HUMAID$_{merge}$. Thus for each dataset, an experiment consists in the average of n runs, each run with $n-1$ crisis types for training and the remaining crisis type for testing. For HUMAID$_{merge}$, as a random data sampling is done so that the dataset can be comparable with the distribution of the French one and as the initial English dataset is pretty large (around 86K), we created 5 independent batches of 16,202 random tweets per batch conserving the same class distribution as KOZLOWSKI$_{merge}$ and reported the averaged macro F-scores.

For both datasets, we replaced all numbers and URLs by a dedicated token and removed user mentions.

4.2 Methods

In addition to our super-resolution data augmentation method, we evaluate several state-of-the-art approaches:

4.2.1 Data-based approaches

- **Synonym replacement** (Zhang et al., 2015): This method substitutes a random word in a text by one of his synonym selected in a thesaurus based on a language model prediction. We used the implementation from the NLPAUG library.[4]

- **Contextual augmentation**: Based on Kobayashi (2018), this method requires a deep learning model calculating the word probability at a given position based on its context. Then, it replaces the word at this position by the predicted word.

- **Back Translation** (Yu et al., 2018): This method translates a text from a source language

[3] https://en.wikipedia.org/wiki/Category:Natural_disasters_by_year

[4] https://github.com/makcedward/nlpaug

Text	Method
(1) #Harvey has so far cost at least 31 lives, forced 30,000+ people from homes in TX and caused extensive destruction.	Original tweet
(2) #Harvey has cost at least 31 lives so far, forced more than 30,000 people to leave their homes in TX and caused considerable destruction.	Back translation
(3) # Moreve It has so far cost at quigo 31 liveds, forced 3 et+ over from homes in TX that causes some extensive destruction.	Synonym replacement
(4) Old #Harovey the has so was far we cost at high leostast 1 31 lives, forced 30,000+ people from homes in TX and caused so extensive place destruction.	Contextual augmentation
(5) The impact of #Harvey has been devastating, with 31 lives lost, over 30,000 people displaced from their homes in TX, and widespread destruction. Let's stand together and support those affected by this crisis. #SupportForHarveyVictims #RebuildingTogether	Generative augmentation
(6) Approximately 32,000 people were displaced in shelters across the state by August 31. The George R. Brown Convention Center, the state's largest shelter, reached capacity with 8,000 evacuees. The NRG Center opened as a large public shelter accordingly. More than 210,000 people registered with FEMA for disaster assistance.	**Super-resolution**

Table 2: Examples of augmented data from one tweet.

into a target language and then translates it back to the source. We used Helsinki-NLP/OPUS-MT-train (Tiedemann and Thottingal, 2020)[5] to translate the French dataset into English and then back into French. For the English dataset, we translated it into French and then back into English.

- **Generative augmentation**: Inspired by Yoo et al. (2021), we used GPT3 (Brown et al., 2020) to generate new data similar to tweets of our datasets. The prompt asks to generate 10 new tweets similar to 5 tweets from the original dataset that share the same label but can be from different crises.

To be able to make comparison between classification models, we used the above methods to augment data with the same distribution as FR$_{AUTO\_NO\_SOCIAL}$ to augment KOZLOWSKI$_{merge}$ and EN$_{AUTO\_NO\_SOCIAL}$ to augment HUMAID$_{merge}$. Table 2 presents examples of augmented data obtained from one tweet with the above methods, showing for some of them the low quality of what is generated.

4.2.2 Model-based approaches

We also compare with the following two model-based methods as they have been shown to be quite effective for crisis management.

- **Manifold Mixup** (Verma et al., 2019): It has been adapted to crisis management by Chowdhury et al. (2020a). It makes semantic interpolation during training, obtaining neural networks (here, FlauBERT$_{Fine-Tuned}$ for French and RoBERTa for English) with smoother decision boundaries.[6]

- **Multi-task learning** (Ye et al., 2019): This strategy trains several layers on different but related tasks to improve the performance. This learning architecture has already been used for crisis management by Kozlowski et al. (2020) and Khanal and Caragea (2021). Following Kozlowski et al. (2020), the tasks are utility (useful vs. not useful), urgency (urgent vs. not urgent vs. not useful) and intention (5 classes) detection, the latter being our main task. For all tasks, we used the Focal loss function and **FlauBERT**$_{Fine-Tuned}$ for French and **RoBERTa** for English.

5 Results

5.1 Quantitative results

Table 3 presents the results obtained by FlauBERT$_{Fine-Tuned}$ on the KOZLOWSKI$_{merge}$ dataset (5 classes) and its augmented versions. The model obtains the best results on the dataset augmented with **FR**$_{AUTO\_NO\_SOCIAL}$ via the automatic super-resolution method. We note that augmenting only the classes DAMAGE and ADVICE-WARNING with few high-resolution texts from Wikipedia can even help improving the performance on the class SOCIAL which has not been augmented. Indeed, Wikipedia contains informative data, similar to what is found in tweets for both classes DAMAGE and ADVICE-WARNING whereas it contains almost no SOCIAL data. Thus, augmenting only the first two classes helps the model to be more discriminant on the class SOCIAL.

Table 4 presents the results obtained by RoBERTa on the HUMAID$_{merge}$ dataset (5 classes)

[5] https://github.com/Helsinki-NLP/OPUS-MT-train

[6] https://github.com/vikasverma1077/manifold_mixup

Dataset	Damage	Advice_Warning	Social	Other	Not Useful	Total
Kozlowski$_{merge}$	54.87	46.94	54.62	23.86	82.83	52.62
Manifold Mixup	55.95	46.67	50.65	26.00	82.64	52.38
Multi-task	57.76	47.15	50.99	**30.50**	82.86	53.85
Kozlowski$_{merge}$ +Back Translation	48.94	46.48	49.25	26.40	**83.00**	50.81
Kozlowski$_{merge}$ +Generative augmentation	52.77	47.66	50.74	22.09	81.83	51.02
Kozlowski$_{merge}$ +Contextual augmentation	58.37	48.04	50.84	24.41	82.05	52.74
Kozlowski$_{merge}$ +Synonym replacement	57.04	47.38	51.35	**28.12**	81.22	53.02
FR$_{AUTO}$	55.67	47.75	50.18	22.51	81.37	51.50
FR$_{AUTO\_NO\_SOCIAL}$	**58.65**	**48.33**	**56.00**	25.73	82.26	**54.19**

Table 3: Results of FlauBERT$_{Fine-Tuned}$ in terms of average F-score on Kozlowski$_{merge}$ for *out-of-type* experiment.

Dataset	Damage	Advice_Warning	Social	Other	Not Useful	Total
HumAID$_{merge}$ (5 batches)	69.87	73.04	82.51	39.18	92.05	71.33
Manifold Mixup	70.42	72.78	82.30	41.26	**92.26**	71.81
Multi-task	73.99	68.04	81.16	37.43	91.47	70.42
HumAID$_{merge}$ +Back Translation	74.86	72.66	81.21	37.31	90.77	71.36
HumAID$_{merge}$ +Generative augmentation	68.46	71.14	75.89	31.26	91.12	67.57
HumAID$_{merge}$ +Contextual augmentation	87.59	71.11	82.06	41.78	91.56	74.82
HumAID$_{merge}$ +Synonym replacement	**87.61**	**73.50**	81.17	41.03	91.78	**75.02**
EN$_{AUTO\_NO\_SOCIAL}$	84.90	73.02	**83.52**	**42.03**	91.38	74.97

Table 4: Results of RoBERTa in terms of average F-score on HumAID$_{merge}$ for *out-of-type* experiment.

Dataset	Damage	Advice_Warning	Social	Other	Not Useful	Total
HumAID$_{merge}$ (5 batches)	71.23	65.58	93.66	42.31	80.48	70.65
Manifold Mixup	70.79	67.12	93.57	42.58	80.12	70.83
Multi-task	73.39	68.49	78.63	40.35	93.87	70.95
HumAID$_{merge}$ +Back Translation	70.99	72.37	93.49	41.76	79.13	71.54
HumAID$_{merge}$ +Generative augmentation	69.77	64.63	93.88	33.68	72.05	66.80
HumAID$_{merge}$ +Contextual augmentation	69.33	86.84	93.44	45.35	80.00	74.99
HumAID$_{merge}$ +Synonym replacement	**72.63**	86.14	93.46	44.65	78.48	75.07
EN$_{AUTO\_NO\_SOCIAL}$	71.52	**86.99**	**93.88**	**45.74**	**81.70**	**75.97**

Table 5: Results of RoBERTa in terms of average F-score on HumAID$_{merge}$ for *out-of-type* experiment without the test run on Earthquake.

and on its augmented versions. Except *Generative augmentation* and *Multi-task*, all augmentation methods outperform the baseline on the original dataset. We can make a similar observation as for the French dataset: augmenting only the class Damage with EN$_{AUTO\_NO\_SOCIAL}$ helps improving the performance on the class Social. SRL4NLP is second behind *Synonym replacement* and its results are statistically significant compared to those obtained on the original HumAID$_{merge}$ dataset (*p-value*=1.68e-12 with the McNemar test). When looking into details at the 5 runs on each crisis type, we noticed that Earthquake was an outlier for which our model underperformed. The test run on Earthquake has a F-score of 70.99 when augmented with SRL4NLP whereas the F-score is 74.05 on the original dataset: this crisis type has the strongest decrease compared to other crisis types. This may be explained by the fact that only 57 paragraphs have been added to augment the data for earthquake, representing much less data compared to other crisis types (at least 100 paragraphs). Table 5 presents the results obtained without the test run on earthquake. In this configuration, SRL4NLP outperforms all state-of-the-art augmentation methods and is more efficient for all classes except Damage.

Concerning recall and precision, FlauBERT$_{Fine-Tuned}$ obtained a precision of 55.24 (resp. 55.98) and a recall of 53.33 (resp. 55.53) on Kozlowski$_{merge}$ (resp. when augmented with SRL4NLP). RoBERTa obtained a precision of 72.01 (resp. 74.03) and a recall of 73.11 (resp. 77.85) on HumAID$_{merge}$ (resp. when augmented with SRL4NLP). SRL4NLP increases mainly recall, which fits particularly with the needs of crisis management where the emergency services need to be aware of all urgent messages.

5.2 Error analysis

A manual error analysis highlights three main causes of misclassification. For example, in Table 6, the predicted label in messages (1) and (2) seems more accurate (support) than the gold label. Indeed, the tweet (1) is about displaced people so it can be considered as DAMAGE but the main objective of the tweet is an intent to support these people. In tweet (2), the support is represented by the hashtag *#PrayForGreece*. In tweet (3), the word *damage* misled the model to consider this tweet as DAMAGE. Finally, message (4) could have multiple labels: the gold label ADVICE-WARNING since the tweet declares a state of emergency but also the predicted one DAMAGE since it mentions people evacuation.

6 Related Work

6.1 Data Augmentation in NLP

Various methods have been developed to generate artificial data in order to enhance the size and the quality of training datasets, avoiding collecting and annotating data manually (Feng et al., 2021). Language models can be used for back translation (Yu et al., 2018) or for synonym replacement and contextual insertion (Zhang et al., 2015; Kobayashi, 2018), keeping the semantic of the original sentences in the artificial data. Coulombe (2018) proposed error injection to be more robust to the noise found in social media posts. Rule-based methods can also be used to add, replace or swap words in a sentence (Wei and Zou, 2019), avoiding using costly language models. Recently, generative language models such as GPT3 have also been employed to generate data and augment existing datasets in a few/zero-shot scenario by using the class label and dedicated prompts as cue for the model (Anaby-Tavor et al., 2020; Yoo et al., 2021; Chia et al., 2022; Ocampo et al., 2023).

Model-based approaches on the other hand make use of multi-task learning leveraging auxiliary tasks (Spangher et al., 2021), knowledge distillation that trains smaller model with soft labels (Tan et al., 2022), and text interpolation that mixes up original and perturbed samples in the hidden space (Chen et al., 2022).

In this paper, we relied on advances in data augmentation for image classification and propose SRL4NLP, a novel augmentation paradigm adapting the well-established super-resolution learning to text classification. Our approach allows for adding new linguistic patterns in the train set relying on structured external textual sources and avoids generating any new data by using existing high-resolution data for a low computational cost.

6.2 NLP for Crisis Management

Social media, such as Twitter, can give precious real-time information in crisis situations (Reuter et al., 2018) (e.g., more than 1,000,000 tweets posted during in Türkiye and Syria earthquake in 2023 (Toraman et al., 2023)), making NLP-based crisis management a hot research topic. Recently, shared tasks have proposed, such TREC-IS (Incident Streams track)[7] (McCreadie et al., 2019, 2020), and CrisisFACTS2022[8] (McCreadie and Buntain, 2023) which allows the development of classifiers in three settings: (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) intents (damage reports, warnings, critics, etc.).

Classifiers are trained in a supervised way with either traditional feature-based learning algorithms (Alam et al., 2021; Li et al., 2018; Kaufhold et al., 2020) or deep learning architectures (Caragea et al., 2016; Neppalli et al., 2018; Kersten et al., 2019; Kozlowski et al., 2020; Chowdhury et al., 2020b; Liu et al., 2021; Wang et al., 2021; Dusart et al., 2021). Overall, results show that intent detection is the most challenging task due to the extremely imbalanced nature of crisis datasets: urgent messages are minority (e.g., about 3.87% and 1.93% for infrastructure/human damages respectively in CrisisFACTS).

The imbalanced issue has however received less attention in the literature. Among existing works, Bayer et al. (2021) used text generation, Kruspe et al. (2018) relied on automatic round-trip translation augmenting under-represented classes in the TREC-IS dataset. Ramachandran and Ramasubramanian (2021) proposed a synonym and hypernym replacement strategy based on WordNet. Finally Chowdhury et al. (2020a) explored text interpolation via Manifold Mixup in a multilingual setting

[7] https://www.dcs.gla.ac.uk/~richardm/TREC_IS/
[8] https://crisisfacts.github.io/

Text	Gold label	Predicted Label
(1) Want to help support the thousands of residents displaced by the FortMcMurray massive wildfire Don t get scammed	Damage	Social
(2) Athens is either under wildfires or under storms i hate this summer PrayForGreece	Not Useful	Social
(3) Fort McMurray Trudeau arrives to view wildfire damage	Other	Damage
(4) Alberta Canada declared a state of emergency after a wildfire forced more than numero numero people to evacuate	Advice_Warning	Damage

Table 6: Examples of misclassification by RoBERTa on HUMAID$_{merge}$ dataset.

with BERT. The use of super-resolution in the field is novel.

7 Conclusion

By drawing an analogy between image and text processing, we have proposed SRL4NLP, a new approach based on super-resolution learning to augment low-resolution imbalanced training data (for example, tweets) with similar texts having a higher resolution (paragraphs).

We have evaluated its effectiveness on urgency detection from tweets posted in crisis situations, relying on imbalanced benchmark datasets in both English and French. We have compared the classification performances over these datasets augmented with few data obtained with several data augmentation methods. The results show statistically significant F-scores in out-of-type evaluation setting, outperforming existing data-based and model-based augmentation approaches. SRL4NLP has also a very low computational cost since it avoids the generation of artificial data and their manual annotation.

For future work, we plan to rely on other external sources for augmenting our data (e.g., news, government reports, return of operating experiences) as well as experiment with other use cases (see below).

Acknowledgements

The research of Farah Benamara is partially supported by DesCartes: The National Research Foundation, Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) program.

Limitations

In this paper, we have evaluated SRL4NLP on crisis management benchmark datasets in English and French but the method and the evaluation protocol for this use case can be extended to datasets in other languages since Wikipedia pages are available in many languages.

We think that it can be also easily transfered to other classification use cases for which low-resolution datasets and high-resolution comparable external resources are available: for example, hate speech detection by augmenting datasets of tweets with posts on discussion forums; political opinion detection by augmenting datasets from social media with political speeches; medical diagnosis by augmenting patient records with medical publications, etc.

Ethics Statement

The data that was used for conducting the experiments is composed of texts from the public domain taken from either datasets publicly available to the research community, or Wikipedia. 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.

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.

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.

Ateret Anaby-Tavor, Boaz Carmeli, Amir Kantor, George Kour, Segev Shlomov, Naama Tepper, and Naama Zwerdling. 2020. Do not have enough data? deep learning to the rescue! In *AAAI*, pages 7383–7390. AAAI Press.

Gunjan Ansari, Muskan Garg, and Chandni Saxena. 2021. Data augmentation for mental health classification on social media. In *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, pages 152–161, National Institute of Technology Silchar, Silchar, India. NLP Association of India (NLPAI).

Markus Bayer, Marc-André Kaufhold, Björn Buchhold, Marcel Keller, Jörg Dallmeyer, and Christian Reuter. 2021. Data augmentation in natural language processing: A novel text generation approach for long and short text classifiers. *CoRR*, abs/2103.14453.

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)*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Mateusz Buda, Atsuto Maki, and Maciej A. Mazurowski. 2018. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259.

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.

Hui Chen, Wei Han, Diyi Yang, and Soujanya Poria. 2022. DoubleMix: Simple interpolation-based data augmentation for text classification. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4622–4632, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Yew Ken Chia, Lidong Bing, Soujanya Poria, and Luo Si. 2022. RelationPrompt: Leveraging prompts to generate synthetic data for zero-shot relation triplet extraction. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 45–57, Dublin, Ireland. Association for Computational Linguistics.

Jishnu Chowdhury, Cornelia Caragea, and Doina Caragea. 2020a. Cross-lingual disaster-related multi-label tweet classification with manifold mixup. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 292–298, Online. Association for Computational Linguistics.

Jishnu Ray Chowdhury, Cornelia Caragea, and Doina Caragea. 2020b. On identifying hashtags in disaster twitter data. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI)*.

Claude Coulombe. 2018. Text data augmentation made simple by leveraging nlp cloud apis. *ArXiv*, abs/1812.04718.

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.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988, Online. Association for Computational Linguistics.

Kai Fukami, Koji Fukagata, and Kunihiko Taira. 2023. Super-resolution analysis via machine learning: a survey for fluid flows. *Theoretical and Computational Fluid Dynamics*, pages 1–24.

Tianyu Gao, Xu Han, Hao Zhu, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. 2019. FewRel 2.0: Towards more challenging few-shot relation classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6250–6255, Hong Kong, China. Association for Computational Linguistics.

Sophie Henning, William Beluch, Alexander Fraser, and Annemarie Friedrich. 2023. A survey of methods for addressing class imbalance in deep-learning based natural language processing. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 523–540, Dubrovnik, Croatia. Association for Computational Linguistics.

Muhammad Imran, Shady Mamoon Elbassuoni, Carlos Castillo, Fernando Diaz, and Patrick Meier. 2013.

Extracting information nuggets from disaster-related messages in social media. *Proc. of ISCRAM, Baden-Baden, Germany*.

Gunilla Elleholm Jensen. 2012. *Key criteria for information quality in the use of online social media for emergency management in New Zealand*. Ph.D. thesis, Open Access Te Herenga Waka-Victoria University of Wellington.

Ehsan Kamalloo, Mehdi Rezagholizadeh, and Ali Ghodsi. 2022. When chosen wisely, more data is what you need: A universal sample-efficient strategy for data augmentation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1048–1062, Dublin, Ireland. Association for Computational Linguistics.

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.

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*.

Sarthak Khanal and Doina Caragea. 2021. Multi-task learning to enable location mention identification in the early hours of a crisis event. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4051–4056, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sosuke Kobayashi. 2018. Contextual augmentation: Data augmentation by words with paradigmatic relations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 452–457, New Orleans, Louisiana. Association for Computational Linguistics.

Rani Koshy and Sivasankar Elango. 2022. Multimodal tweet classification in disaster response systems using transformer-based bidirectional attention model. *Neural Computing and Applications*, 35.

Michał Koziarski and Bogusław Cyganek. 2018. Impact of low resolution on image recognition with deep neural networks: An experimental study. *International Journal of Applied Mathematics and Computer Science*, 28(4):735–744.

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):102284.

Anna Kruspe, Jens Kersten, Matti Wiegmann, Benno Stein, and Friederike Klan. 2018. Classification of incident-related tweets: Tackling imbalanced training data using hybrid cnns and translation-based data augmentation. In *Text REtrieval COnference (TREC)*.

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.

Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li. 2020. Dice loss for data-imbalanced NLP tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 465–476, Online. Association for Computational Linguistics.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2020. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):318–327.

Bo Liu, Jiupeng Tang, Haibo Huang, and Xi-Yun Lu. 2020. Deep learning methods for super-resolution reconstruction of turbulent flows. *Physics of Fluids*, 32(2). 025105.

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.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

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*.

Tobias Mayer, Santiago Marro, Elena Cabrio, and Serena Villata. 2021. Enhancing evidence-based medicine with natural language argumentative analysis of clinical trials. *Artificial Intelligence in Medicine*, 118:102098.

Richard McCreadie and Cody Buntain. 2023. Crisisfacts: Building and evaluating crisis timelines.

Richard McCreadie, Cody Buntain, and Ian Soboroff. 2019. TREC Incident Streams: Finding Actionable Information on Social Media. In *Proceedings of the 16th International Conference on Information Systems for Crisis Response and Management, València, Spain, May 19-22, 2019*. ISCRAM Association.

Richard McCreadie, Cody Buntain, and Ian Soboroff. 2020. Incident streams 2019: Actionable insights and how to find them. In *Proceedings of the 17th ISCRAM Conference*.

Elias Nehme, Lucien E. Weiss, Tomer Michaeli, and Yoav Shechtman. 2018. Deep-storm: super-resolution single-molecule microscopy by deep learning. *Optica*, 5(4):458–464.

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.

Rajvardhan Oak, Min Du, David Yan, Harshvardhan Takawale, and Idan Amit. 2019. Malware detection on highly imbalanced data through sequence modeling. In *Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security*, AISec'19, page 37–48, New York, NY, USA. Association for Computing Machinery.

Nicolas Ocampo, Elena Cabrio, and Serena Villata. 2023. Playing the Part of the Sharp Bully: Generating Adversarial Examples for Implicit Hate Speech Detection. In .

Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. 2014. Crisislex: A lexicon for collecting and filtering microblogged communications in crises. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1):376–385.

Alexandra Olteanu, Sarah Vieweg, and Carlos Castillo. 2015. What to expect when the unexpected happens: Social media communications across crises. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '15, page 994–1009, New York, NY, USA. Association for Computing Machinery.

Xingchao Peng, Judy Hoffman, Stella X. Yu, and Kate Saenko. 2016. Fine-to-coarse knowledge transfer for low-res image classification. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 3683–3687.

Dharini Ramachandran and Parvathi Ramasubramanian. 2021. A novel domain and event adaptive tweet augmentation approach for enhancing the classification of crisis related tweets. *Data Knowl. Eng.*, 135:101913.

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.

Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI EA '21, New York, NY, USA. Association for Computing Machinery.

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.

Cinthia Sánchez, Hernan Sarmiento, Andres Abeliuk, Jorge Pérez, and Barbara Poblete. 2023. Cross-lingual and cross-domain crisis classification for low-resource scenarios. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, pages 754–765.

Efsun Sarioglu Kayi, Linyong Nan, Bohan Qu, Mona Diab, and Kathleen McKeown. 2020. Detecting urgency status of crisis tweets: A transfer learning approach for low resource languages. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4693–4703, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Connor Shorten and Taghi Khoshgoftaar. 2019. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6.

Alexander Spangher, Jonathan May, Sz-Rung Shiang, and Lingjia Deng. 2021. Multitask semi-supervised learning for class-imbalanced discourse classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 498–517, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Qingyu Tan, Ruidan He, Lidong Bing, and Hwee Tou Ng. 2022. Document-level relation extraction with adaptive focal loss and knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1672–1681, Dublin, Ireland. Association for Computational Linguistics.

Rui Tian, Zuxuan Wu, Qi Dai, Han Hu, Yu Qiao, and Yu-Gang Jiang. 2023. ResFormer: Scaling ViTs with Multi-Resolution Training. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

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*.

Ruud J.G. van Sloun, Oren Solomon, Matthew Bruce, Zin Z Khaing, Yonina C. Eldar, and Massimo Mischi. 2019. Deep learning for super-resolution vascular ultrasound imaging. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1055–1059.

Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. 2019. Manifold mixup: Better representations by interpolating hidden states. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6438–6447. PMLR.

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*.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Wenming Yang, Xuechen Zhang, Yapeng Tian, Wei Wang, Jing-Hao Xue, and Qingmin Liao. 2019. Deep learning for single image super-resolution: A brief review. *Trans. Multi.*, 21(12):3106–3121.

Wei Ye, Bo Li, Rui Xie, Zhonghao Sheng, Long Chen, and Shikun Zhang. 2019. Exploiting entity BIO tag embeddings and multi-task learning for relation extraction with imbalanced data. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1351–1360, Florence, Italy. Association for Computational Linguistics.

Kang Min Yoo, Dongju Park, Jaewook Kang, Sang-Woo Lee, and Woomyoung Park. 2021. GPT3Mix: Leveraging large-scale language models for text augmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2225–2239, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mustafa Z Yousif, Linqi Yu, and Hee-Chang Lim. 2021. High-fidelity reconstruction of turbulent flow from spatially limited data using enhanced super-resolution generative adversarial network. *Physics of Fluids*, 33(12):125119.

Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. 2018. Qanet: Combining local convolution with global self-attention for reading comprehension. *arXiv preprint arXiv:1804.09541*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Wilman W. W. Zou and Pong C. Yuen. 2012. Very low resolution face recognition problem. *IEEE Transactions on Image Processing*, 21(1):327–340.

A French Lexicon

Damage: Bilan, Victimes, Répercussion, Dommage, Impact, Conséquences, Dégâts, Infrastructure, Humain, Panne, Séquelle, Pertes, Rescapés, Dévastation, Effet, Destruction

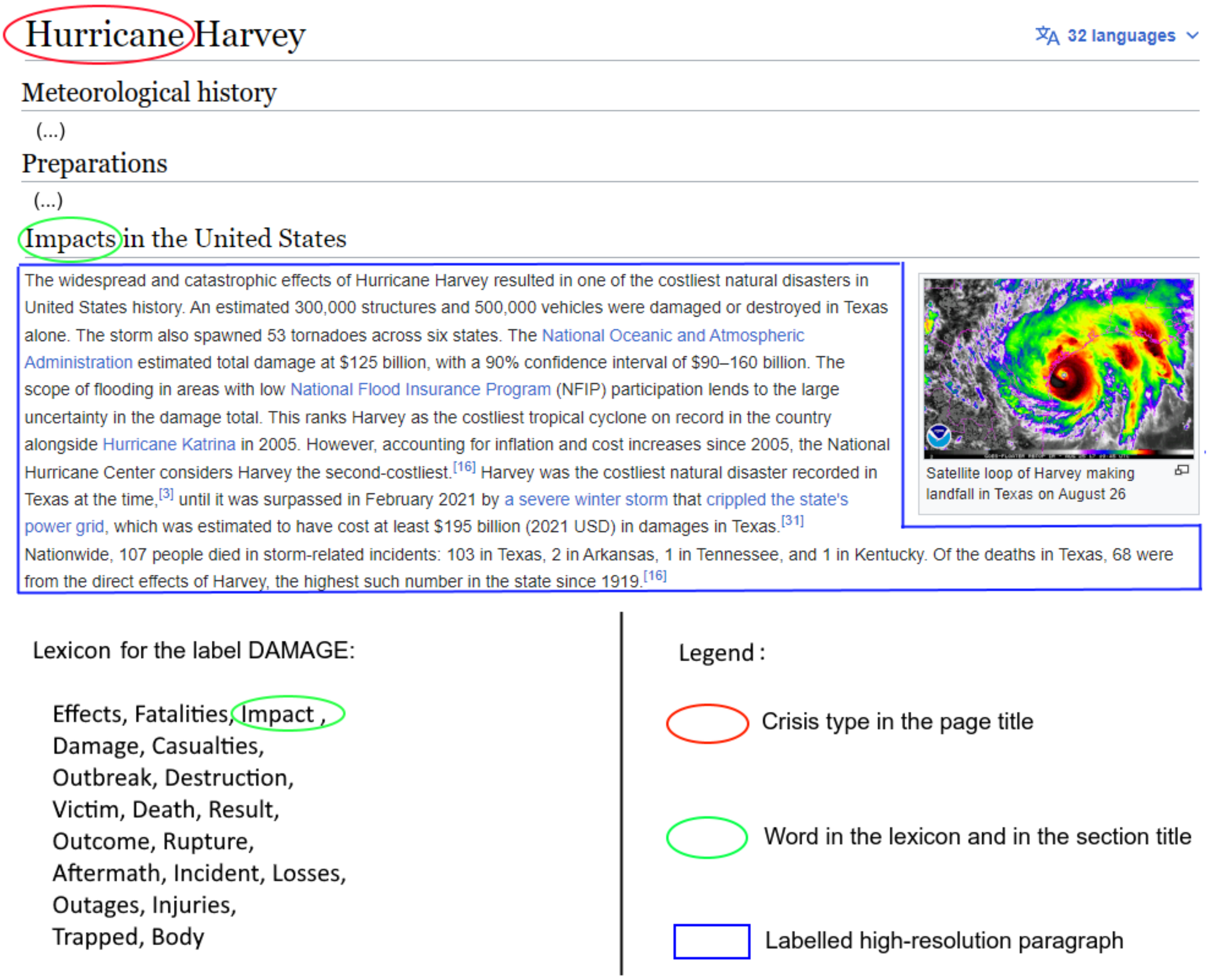
Advice_Warning: Précédant, Repère, Chronologie, Précurseur, Préparatifs, Contexte, Situation, Préparation, Prémice, Prévision, Évolution, Plan

Social: Réactions, Secours, Reconstruction, Presse, Communiqué, Judiciaire, Controverse, Complotiste, Instrumentalisation, Solidarité, Sociale, Aide, Restauration, Hommage, Commémoration, Communauté

B English Lexicon

Damage: Effects, Fatalities, Impact ,Damage, Casualties, Outbreak, Destruction, Victim, Death, Result, Outcome, Rupture, Aftermath, Incident, Losses, Outages, Injuries, Trapped, Body

C Scrapper and Automatic Annotation



Hurricane Harvey

Meteorological history

(...)

Preparations

(...)

Impacts in the United States

The widespread and catastrophic effects of Hurricane Harvey resulted in one of the costliest natural disasters in United States history. An estimated 300,000 structures and 500,000 vehicles were damaged or destroyed in Texas alone. The storm also spawned 53 tornadoes across six states. The National Oceanic and Atmospheric Administration estimated total damage at $125 billion, with a 90% confidence interval of $90–160 billion. The scope of flooding in areas with low National Flood Insurance Program (NFIP) participation lends to the large uncertainty in the damage total. This ranks Harvey as the costliest tropical cyclone on record in the country alongside Hurricane Katrina in 2005. However, accounting for inflation and cost increases since 2005, the National Hurricane Center considers Harvey the second-costliest.[16] Harvey was the costliest natural disaster recorded in Texas at the time,[3] until it was surpassed in February 2021 by a severe winter storm that crippled the state's power grid, which was estimated to have cost at least $195 billion (2021 USD) in damages in Texas.[31] Nationwide, 107 people died in storm-related incidents: 103 in Texas, 2 in Arkansas, 1 in Tennessee, and 1 in Kentucky. Of the deaths in Texas, 68 were from the direct effects of Harvey, the highest such number in the state since 1919.[16]

Satellite loop of Harvey making landfall in Texas on August 26

Lexicon for the label DAMAGE:

Effects, Fatalities, Impact ,
Damage, Casualties,
Outbreak, Destruction,
Victim, Death, Result,
Outcome, Rupture,
Aftermath, Incident, Losses,
Outages, Injuries,
Trapped, Body

Legend :

○ Crisis type in the page title

○ Word in the lexicon and in the section title

□ Labelled high-resolution paragraph

Figure 2: Example of the automatic annotation of Wikipedia paragraphs (https://en.wikipedia.org/wiki/Hurricane_Harvey).