Noname manuscript No.
(will be inserted by the editor)

The usefulness of NLP techniques for predicting peaks in firefighter interventions due to rare events

Selene Cerna · Christophe Guyeux · David Laiymani*

the date of receipt and acceptance should be inserted later

Abstract blue In some countries such as France, the number of operations assisted by firefighters has shown an almost linear increase over the years, contrary to their resource capacity. For this reason, predicting the number of interventions has become a necessity. Initially, time series models were developed with several types of qualitative and quantitative features, including the alert level of the bulletins, to predict the operational load. We realized that interventions related to human activities are quite predictable. However, the recognition of interventions due to rare events such as storms or floods needs more than quantitative meteorological data to be identified, since there are almost always zero cases. Thus, this work proposes the application of Natural Language Processing (NLP) techniques, namely, Long Short-Term Memory, Convolutional Neural Networks, FlauBERT, and CamemBERT to extract features from the texts of weather bulletins in order to recognize periods with peak interventions, where the intense workload of firefighters is caused by rare events. Four categories identified as Emergency Person Rescue, Total Person Rescue, interventions related to Heating, and Storm/Flood were our targets for the multilabel classification models developed. The results showed a remarkable accuracy of 80%, 86%, 92%, and 86% for Emergency Rescue People, Total Rescue People, Heating, and Storm/Flood, respectively.

Keywords Firemen activity prediction, Regression, Natural language processing, BERT

1 Introduction

The missions and the activity of fire departments change from one country to another: in some countries, fire departments are only in charge of extinguishing fires, while in others they are also in charge of rescuing people, whether it is urgent or not. In countries such as France, fire activities represent only almost 10% of all their interventions, and they may be called out for road accidents, floods, respiratory ailments, gastroenteritis, or even wasp nests. In Western countries with an aging population, the share of personal assistance is constantly increasing, when repeated economic crises lead to budget cuts.

Thus, in countries where fire brigades activity encompass rescuing people, the number of interventions has been steadily increasing for several years now and the Covid-19 pandemic has only amplified the situation. In this context, ensuring rapid and efficient interventions becomes a major challenge for many brigades and it seems interesting to try to plan interventions in advance. Better still, answering the questions When? Where? Of what type (road accident, domestic accident, flood...)? can have a strong impact and therefore save lives. This difficult problem is just beginning to be studied [1–3]. In fact, the vast majority of brigades have accumulated an important mass of data related to their past interventions. It is therefore interesting to try to use these data and the recent advances in Machine Learning (ML) to try out intervention prediction. It is reasonable to think that such predictions are possible, as the reasons for these interventions are to some extent

[*] Authors have contributed equally to this research work and David Laiymani is the corresponding author.

Selene Cerna · Christophe Guyeux · David Laiymani
Femto-ST Institute, UMR 6174 CNRS
Univ. Bourgogne Franche-Comté
Belfort, France
E-mail: firstName.name@univ-fcomte.fr

deterministic. Forest fires occur more frequently in dry, hot weather than in wet, rainy weather; floods follow heavy rains; domestic and work accidents occur mostly in the middle of the day, and they are rare at 3am ; falls on ice do not occur in summer and drowning in outdoor pools does not occur in winter. As can be seen from the above examples, weather conditions have an undeniable impact on human activity and the accidents it causes, and therefore on the activity of firefighters. In addition, most national weather prediction services provide online services or APIs to retrieve various physical quantities useful for weather knowledge. These quantities, which are internationally codified and are mandatory measurements, include temperature, pressure, wind direction and strength, dew point, or hydrometry, for a set of stations spread over the entire territory.

However, while these physical quantities are useful in predicting, to some extent, the weather, they are only partially effective in predicting firefighters interventions, their types and intensities. Knowing that a thunderstorm event is coming does not accurately predict the extent to which firefighters activity will be affected. More precisely, by coupling the history of weather conditions with the history of interventions, we can see that while some storm-type weather conditions clearly lead to peaks in intervention, others, on the contrary, go somewhat unnoticed. In other words, the simple quantitative value of the physical quantities to be measured does not alone make it possible to accurately predict the occurrence of periods of heavy intervention (for example, heavy rains that do not systematically lead to flooding). In this article, we would like to insist on the fact that most states have also set up a meteorological vigilance service, with reports indicating the risk incurred (snow, ice, storm, floods...), the duration of the event and its textual description, as well as useful advice. It is clear that integrating the level of vigilance in the form of qualitative variables, one for each risk monitored, allows on the one hand to improve the regression scores in the learning phase of the total number of interventions. On the other hand, we argue that the textual content of newsletters is valuable and under-exploited, that sentences such as "it is advisable to unplug electrical appliances" or "it is strongly advised not to walk along the coast" are rich in information for the prediction of interventions such as heating or emergency rescue due to rare events, and that automatic natural language processing tools are now mature enough to understand the link between such assertions and peak intervention periods.

blue

Purpose and Contributions

With these elements in mind, the present work performs a detailed study of various ML models developed for the prediction of intervention peaks due to rare weather events and identifies the remarkable impact of NLP models and weather bulletin texts. The rarity of events such as flood or heating events makes them difficult to predict due to the the small amount of data available and, as we will see in the remainder, these events can be the source of extremely high numbers of intervention. The predictions are made for 4 categories of interventions, namely, Emergency Person Rescue, Total Person Rescue, interventions related to Heating, and interventions related to Storm/Flood. The data was collected from 2012-2020. They are composed of interventions of the Fire Department of Doubs (SDIS 25), located at the north-east of France, quantitative weather variables and texts of weather bulletins and their vigilance levels from Météo-France. In this way, SDIS 25 or other Emergency Medical Services (EMS), in general, could better recognize periods with high workload generated by rare events, through text processing of weather bulletins, and strategically prepare the appropriate personnel and armament to deal with the crisis so that no service disruptions occur and more lives can be saved. In summary, this article proposes 4 main contributions described as follows.

a. *Analysis and prediction of interventions using basic univariate time series models.* This allows to recognize the seasonality of rescue-type interventions and to make predictions with a small margin of error, but demonstrates the lack of recognition of interventions due to rare events. Here we compared 3 basic models that generated predictions of the number of interventions per hour for the 4 categories independently: equal to the mean, equal to the last known value, and equal to the mean per hour.
b. *Analysis and prediction of interventions using multivariate time series models.* This allows to complement and identify the trend and the seasonality of the signal by qualitative and quantitative variables, such as vigilance indicators and weather and calendar variables, but which are still not sufficient to recognize interventions caused by rare events. Here, several models were developed with the state-of-the-art Extreme Gradient Boosting (XGBoost) technique, in which the features are combined to measure their impact on the prediction of the number of interventions per hour for the 4 categories independently.
c. *Analysis and prediction of intervention peaks using multilabel classification models based on decision trees and tabular data.* The problem is restated as

a multilabel classification task for the 4 categories, using the variables of the previous model and the XGBoost and Random Forest (RF) techniques well-known in the literature for its fast execution and robustness. This allows to determine the influence of the variables according to the new perspective and to deduce that there is still a lack of information to recognize the peaks due to rare events. Thus, these models become our baselines for the following models with Natural Language Processing (NLP).

d. *Analysis and prediction of intervention peaks using multilabel classification models based on NLP techniques and text from meteorological bulletins.* We developed and compared models based on ancient (Long Short-Term Memory – LSTM and Convolutional Neural Network – CNN) and modern (FlauBERT and CamemBERT transformers) NLP techniques and only texts extracted from weather bulletins. This allows to significantly improve the forecast of the peaks compared to the previous models with decision trees and tabular data, and thus demonstrate that it is possible to extract much more information from public weather bulletins using NLP techniques.

The structure of this article continues with the section 2 that reviews the contributions of related works. The section 3 describes the acquisition and preprocessing of the data. The section 3 presents the types of neural networks applied for the natural language processing of bulletins. The section 5 presents the results of 3 basic approaches developed prior to the approach with NLP techniques, to analyze and demonstrate the efficacy of the latter with the best model found. This article ends with the section 6, in which our conclusions are shown and avenues for future improvements are discussed.

2 Related work

Among the works reviewed and related to the optimization of fire departments responses to incidents, we mainly found contributions for the prediction of interventions [3, 4] and fires [5]. Likewise, in certain parts of the world firefighters are part of EMS, since they also provide ambulance services. In this way, predictions of traffic accidents [6], ambulance response time and resource allocation are also included [7–9]. Furthermore, we can find works related to the predictions of rare events such as earthquakes [10] and hurricanes [11], which would allow firefighters to identify a specific location for damage assessment and develop better strategies when succoring the population.

In fact, the aforementioned works did not use NLP techniques. However, in other studies, NLP techniques demonstrate their outstanding utility by enriching data sources and predictions. This is achieved by recovering habits, preferences, emotions, feelings, and distress messages through the recognition of semantic patterns [12–14] from various media such as social networks, the news, therapeutic reminders, information systems. . . under the video-based, audio-based and, text-based formats [15]. On the one hand, in [16], the authors presented a cognitive assistant system for EMSs based on Google Speech API, in which the voice records (incident description and patient status) received by the respondent are converted into texts for extract medical concepts. In this way, the system responds by providing information to rescuers on protocols to follow such as resuscitation and airway management. On the other hand, NLP has contributed to the generation of terminological sources for the classification and forecasting of rare events or crises [17–20]. For example, in [18], it is presented the first study for crisis management using French transformer-based architectures (BERT, FlauBERT, and CamemBERT) apply to French social media, in order to classify tweets for natural disasters. In [21], the authors make use of the Bayesian model averaging approach and linear-chain conditional random fields to extract knowledge from tweets and build a decision support system to identify early warning signs of earthquakes. Also, in [22] is developed the Flood AI Knowledge Engine, which is a system composed of ontology management, query mapping and execution, and NLP modules. The system provides emergency preparedness and response, as well as knowledge about flood-related resources for the population. This knowledge is returned by interpreting natural language queries from users.

Having seen the impact of NLP on the recognition of extreme events and given that, to the authors' knowledge, no previous studies have exploited a source such as weather bulletins to detect trends in the number of firefighter interventions, the present work takes advantage of NLP techniques to process these bulletins and predict peak intervention periods for the categories: Emergency Person Rescue, Total Person Rescue, interventions related to Heating, and Storm/Flood.

3 Data retrieval

3.1 Data acquisition

We collected the interventions of the SDIS25 firefighters over the period 2012-2020, for the entire department, for each time slot of this period (76224 slots, 1 slot = 1

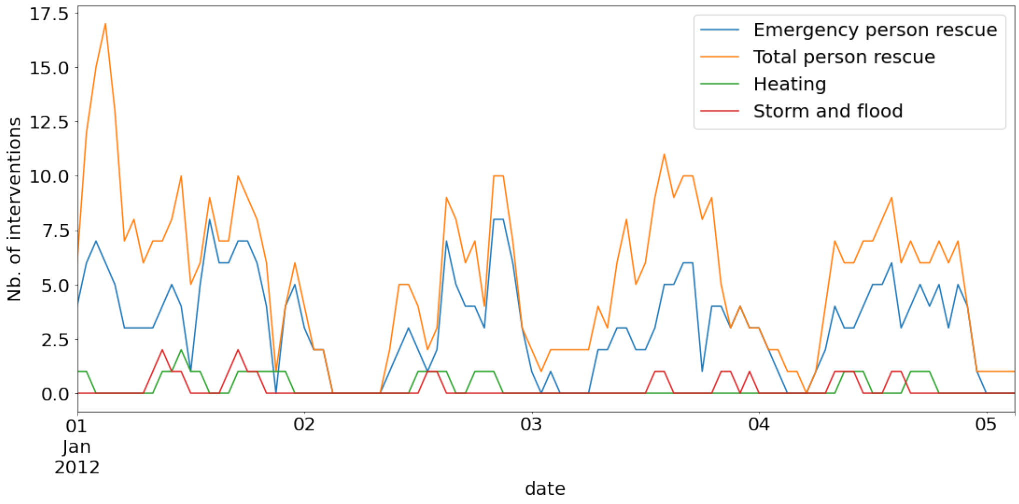


Fig. 1 Interventions for the four types considered, early 2012

hour), and for the following four types of intervention: Emergency Person Rescue, Total Person Rescue, interventions related to Heating, and Storm/Flood. A short statistical description of each type of intervention (per hour) is provided in Table 1, when Figure 1 plots the curves for each type for the first days of 2012.

blue

Table 1 Statistical description of the interventions per hour

Type of intervention	Mean	Std.	Min.	Max
Emergency rescue of people	3.56	2.46	0	20
Total rescue of people	7.25	4.53	0	30
Heating	0.32	0.62	0	6
Storm and Flood	0.26	1.17	0	82

It can be seen that in both types of personal assistance (emergency and total), the number of interventions is on average high enough to show seasonal patterns: the 24-hour cycle is clearly shown in Figure 1. We can also see that some days are special, such as the New Year's Day and that the integration of calendar variables into this daily seasonality should improve the predictions. Finally, the rarity of events such as storms or heating events, will make their predictions problematic in the absence of additional information, which already shows the interest of considering weather-type variables. This is even more true for the ”storms and floods” case as it has the lowest average (0.26 interventions per hour) and an extremely high maximum (82 interventions in one hour): these events are very rare, but the source of an extremely high number of interventions.

This is the reason why we have retrieved historical meteorological data from the Météo-France site (essential SYNOP data [23]). The three closest meteorological stations selected are those of Nancy-Ochey (latitude 48.581000, longitude 5.959833), Dijon-Longvic (47.267833; 5.088333), and Basel-Mulhouse (47.614333; 7.510000). The data recovered in this way are: temperature (degrees Celsius), pressure (Pa), pressure variation (Pa per hour), barometric trend (categorical), humidity (percent), dew point, last hour rainfall (millimeter), last three hours rainfall (millimeter), mean wind speed (10 min., m/s), mean wind direction (10 min., m/s), gusts over a period (m/s), horizontal visibility (m), and current weather (categorical, 100 possible values).

These data were supplemented by (textual) vigilance alert bulletins from Météo-France [24]. They are XML files containing the type of vigilance (heat wave, extreme cold, snow or ice, thunderstorms, strong winds), the beginning and end of the vigilance period, the level of the alert (green, orange or red), a detailed description of the risk (including the locations impacted, the conditions to be expected...), as well as a set of very detailed advice to users. An example of such files is provided in Fig. 2. Finally, we added calendar-type variables, namely the time of the niche considered, the day in the week, the day in the month, the month in the year, and the year considered.

3.2 Data preprocessing

12218 weather bulletins have been produced since 2011, including 1054 for the North-East region (known as CMIRNE) of France, which interests us. 18 departments are represented in this region, but we were only interested in the Doubs and its three neighboring departments (Territory of Belfort, Haute-Saône and Jura) in a first time.

The set of texts available may seem small at first glance, but on the one hand each bulletin covers a number of departments, and has several sections (location,

```
<Titre name="Conséquences␣possibles">
<Paragraphe>
<Intitule>Vent/Orange</Intitule>
<Texte>
* Des coupures d'électricité␣et␣de␣téléphone␣peuvent␣affecter␣les␣réseaux␣de␣distribution
pendant␣des␣durées␣relativement␣importantes.
</Texte>
<Texte>
*␣Les␣toitures␣et␣les␣cheminées␣peuvent␣être␣endommagées.
</Texte>
<Texte>
*␣Des␣branches␣d'arbre risquent de se rompre.- Les véhicules peuvent être déportés.
</Texte>
<Texte>
* La circulation routière peut être perturbée, en particulier sur le réseau secondaire
en zone forestière.
</Texte>
<Texte>
* Quelques dégâts peuvent affecter les réseaux de distribution d'électricité␣et␣de␣téléphone.
</Texte>
```

Fig. 2 Example of a Possible Consequences section (in French)

description, qualification of the phenomenon, new facts, current situation, expected evolution, possible consequences, and behavioural advice), and each section is made up of several long and detailed sentences. The result is a corpus of 76333 characters.

These bulletins are then cut out by section. Over the entire period of vigilance, the average number of interventions is calculated for each of the four types considered, and a class 0 or 1 is introduced depending on whether this number is above the average number of interventions of the type in question, for all the period 2012-2020. We thus associate 4 binary labels for each text in each section of each bulletin, as shown in Table 2. Through this encoding, we became interested in the question: "in the context of this vigilance event, should we expect an increased number of interventions such as heating, or storm, etc.?". In other words, by this increased number, we mean a higher than average number of interventions.

In Figure 3, one can observe in more detail the number of samples (texts) with value 0, where the number of interventions was below or equal to the mean; and value 1, in the opposite case. Furthermore, we find an unequal distribution in each binary class of each category, that might be bias the prediction models.

4 NLP models

Let us now introduce the neural networks that have been considered here for natural language processing.

4.1 LSTM

Long Short-Term Memory (LSTM) networks are a type of Recurrent Neural Network (RNN) which is a category of neural networks dedicated to sequence processing [25]. In the case of NLP, RNN are interesting since, first they are able to process sequences of variable size and second the use of recurrent connections allow to analyze the past part of the signal. In this way RNN are particularly well suited to handle three different types of problems: sequence labeling, sequence classification and sequence generation. A recurrent network can be approximated by a non-recurrent network unfolded in time. But as the unfolded network is deeper, the vanishing of the gradient is more important during learning, and it becomes more difficult to train. In the same way, as the weights of the recurrent layer are duplicated, RNNs are also subject to exploding gradient. Although they are very effective for modeling short- or medium-term dependencies, RNN are still insufficient for modeling long-term or very long-term dependencies. In NLP, it is common to need to model dependencies of the order of a hundred or more time steps. That is why LSTM has been introduced.

In order to model very long-term dependencies, it is necessary to give recurrent neural networks the ability to maintain a state over a long period of time. This is the purpose of LSTM *cells*. The *cell* can be seen as an internal memory and is able to maintain a state for as long as necessary. It consists of a numerical value that the network can control depending on the situation. The memory cell can be controlled by three control gates: the input gate, i_t, decides whether the input

Table 2 Example of a multilabelling of vigilance texts

Text	Emergency Person Rescue	Total Person Rescue	Heating	Storm/ Flood
Les températures sont déjà négatives aujourd'hui mercredi. A 15h les . . . une vigilance particulière notamment pour les personnes sensibles ou exposées.	1	0	1	1
Période de grand froid; moins intense qu'en 1985; mais nécessitant toutefois . . . températures sous abris observées s'échelonnent entre -1 et -4 degrés.	1	0	1	1

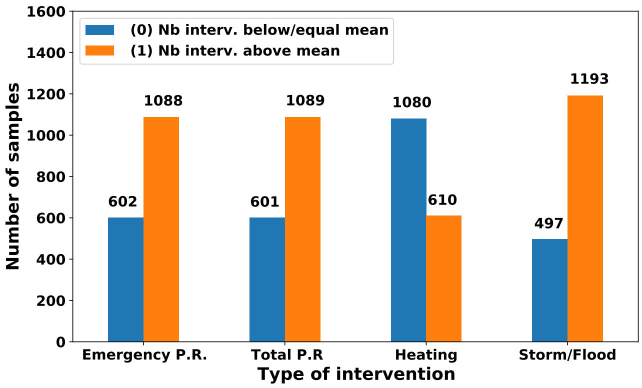


Fig. 3 Number of samples for each category

should modify the content of the cell; the forget gate, f_t, decides which memories will be eliminated from the previous long-term state c_{t-1}; the output gate, o_t, decides whether the cell content should influence the output $y(t)$ of the neuron. Let x_t be the present entry and h_{t-1} the preceding short-term state. LSTM can be expressed in the following way:

$$i_t = \sigma(W_{xi}^T \cdot x_t + W_{hi}^T \cdot h_{t-1} + b_i) \tag{1}$$

$$f_t = \sigma(W_{xf}^T \cdot x_t + W_{hf}^T \cdot h_{t-1} + b_f) \tag{2}$$

$$o_t = \sigma(W_{xo}^T \cdot x_t + W_{ho}^T \cdot h_{t-1} + b_o) \tag{3}$$

$$g_t = \tanh(W_{xg}^T \cdot x_t + W_{hg}^T \cdot h_{t-1} + b_g) \tag{4}$$

$$c_t = f_t \otimes c_{t-1} + i_t \otimes g_t \tag{5}$$

$$y_t = h_t = o_t \otimes \tanh(c_t) \tag{6}$$

where W_{xi}, W_{xf}, W_{xo} and W_{xg} are the weight matrices for their connection to the input vector x_t; W_{hi}, W_{hf}, W_{ho} and W_{hg} are the weight matrices for their connection to the previous short-term state h_{t-1}; and b_f, b_g, b_i and b_o are the bias terms of each layer.

LSTM are generally used in layers and in this case, the outputs of all neurons are fed back to the inputs of all neurons.

blue In this paper, we used the Keras and TensorFlow libraries in Python to initially built several architectures with 1, 2, and 3 LSTM layers, different numbers of neurons, and constant values for learning rate and batch size. Then, we evaluated their performances and selected the 3 architectures that gave us the first best results, these are the ones described in Table 3. Finally, to obtain the best LSTM model, we intensified the search with the 3 selected architectures, varying the learning rate and batch size parameters. For this, we used the HyperOpt library with 100 iterations and the Tree Parzen Estimator suggest (tpe.suggest) algorithm, which models 2 density functions instead of the probability of an observation to estimate the expected improvement of a new configuration.

Furthermore, before the texts enter the neural network, a preprocessing was applied as detailed below:

a. Words with flexible endings were eliminated from the French texts and their base forms were returned, this is known as lemmatization, through the "spacy" library. Unlike a pre-trained model in a large vocabulary, this LSTM model uses the texts of the weather bulletins as a base, it is there the need to bring the words to their base form to identify a greater risk of a possible event. For example, the words "inondations" and "inondables" share the same root and refer to "floods".

Table 3 Defined architectures for LSTM

Archi. 1	Archi. 2	Archi. 3
Input(200)	Input(200)	Input(200)
Embedding(100)	Embedding(100)	Embedding(200)
LSTM(128)	LSTM(1000)	LSTM(128)
Dense(256, ReLU)	Dense(2000, ReLU)	Dense(256, ReLU)
Dropout(0.5)	Dropout(0.5)	Dropout(0.2)
Dense(4, Sigmoid)	Dense(4, Sigmoid)	LSTM(512)
		Dense(1024, ReLU)
		Dropout(0.2)
		Dense(4, Sigmoid)

b. Using the ”nltk” library, French stopwords were eliminated, that is, words such as articles, pronouns, prepositions, auxiliary verbs, among others, which do not have a big impact on our predictions.

c. From the Keras library, we used the Tokenizer function to create a dictionary of words based on their frequency. Then, we reduced the dictionary to the 1000 most repeated words and for each text we generated vectors with integer values, that are the indexes in the dictionary. Finally, they were padded to the same length and entered the neural network described previously.

4.2 CNN

Convolutional neural networks (CNN) is a class of deep neural networks, most often applied to visual image analysis [26]. More recently, however, CNN have also found their place in solving problems related to NLP tasks. A CNN is generally build around the following layers:

- Convolutional layers are composed of neurons whose purpose is to detect patterns (features map) from their inputs.
- Pooling layers whose purpose is to reduce the feature map dimensionality in order to be more computationally efficient. These layers are often chained after convolutional layers.
- These two previous set of layers are generally followed by one or more fully connected layers

As previously stated, CNN has been a game changer in the field of image analysis. They also have shown some interesting results in NLP [18, 27]. Indeed, texts can be represented an array of vectors, just like images can be represented by an array of pixel values. Here, we deal with one dimensional convolutions, but the principles remain the same: we still want find patterns in the sequence which become more complex with each added convolutional layer.

Table 4 Defined architectures for CNN

Archi. 1	Archi. 2	Archi. 3
Input(200)	Input(200)	Input(200)
Embedding(200)	Embedding(200)	Embedding(200)
Conv1D(128, 3, ReLU)	Conv1D(16, 3, ReLU)	Conv1D(256, 3, ReLU)
MaxPooling1D(2)	Dropout(0.2)	Dropout(0.2)
Flatten()	MaxPooling1D(2)	MaxPooling1D(4)
Dense(4, Sigmoid)	Conv1D(32, 3, ReLU)	Conv1D(300, 4, ReLU)
	Dropout(0.5)	Dropout(0.2)
	MaxPooling1D(2)	MaxPooling1D(4)
	Conv1D(64, 3, ReLU)	Conv1D(360, 4, ReLU)
	MaxPooling1D(2)	Dropout(0.5)
	Flatten()	MaxPooling1D(4)
	Dense(4, Sigmoid)	Flatten
		Dense(400, ReLU)
		Dropout(0.2)
		Dense(4, Sigmoid)

In order to associate each word with a specific vector, different techniques can be used. We are talking here about words embedding techniques. The most effective ones are those that are context-sensitive. We can cite here Glove [28] and Word2Vec [29] and as we will see in the next section, Bert is another word embedding technique. Word embeddings are able, by reducing the dimension, to capture the context the semantic and syntactic similarity (gender, synonyms, ...) of a word. For example, one would expect the words ”remarkable” and ”admirable” to be represented by relatively closely spaced vectors in the vector space where these vectors are defined. It has been shown that contextual words embedding can effectively captures the semantic and arithmetic properties of a word. It also reduces the size of the problem and therefore the learning task. In the case of CNN, the obtained inputs vectors will allow the model to have a much better representation of the words during the learning phase.

Given the sequential nature of texts, RNN and LSTM are more common models in NLP. However they can be long and difficult to train. In this way, for large data sets CNN can be an interesting alternative.

blue Similar to the architecture selection procedure performed in LSTM, Table 4 shows the specifications of the three CNN architectures selected to subsequently perform an exhaustive search, varying the learning rate and batch size with the HyperOpt library. In addition, the text preprocessing performed before entering the CNN is also the applied in the LSTM section (4.1).

4.3 Transformers

The sequential nature of RNNs was regularly pointed out as a hindrance to the training of these models on long texts both for computation time reasons (even modern GPUs do not parallelize well this type of process) and because of gradient vanishing problems (despite the use of LSTM models). In order to no longer process texts in a sequential way, Transformers models provides a solution by processing the whole sequence all at once and by using the attention mechanism which allow capture different types of relationships between tokens. A Transformer is built on the basis of an encoder and a decoder, each of them consisting of a stack of attention and dense layers. Since 2017 and the first Transformer model, many models have been developed such as ELMo, GPT-2 and GPT-3, BERT, XLNet, RoBERTa, Turing-NLG... The main advantages of the Transformer model are the following:

- The distance between 2 tokens is no longer a parameter taken into account by the model (the model can take into account long-term dependencies).
- The attention matrix calculation allows to parallelize the process of encoding and then decoding the sequences, thus accelerating the calculations.
- No labeled data are required to pre-train these models and it is then possible to train a transformer-based model by providing a huge amount of unlabeled text data.

From the last point it follows that it is possible to do transfer learning with this trained model in order to perform other NLP tasks like text classification, named entity recognition, text generation... This is how in June 2017, Google presents BERT (Bidirectional Encoder Representations from Transformers) [30]. It is a Transformer composed of a suite of encoders only (N = 12 or 24 depending on the version: base with 110 millions parameters or large with 340 millions parameters). Bert was originally pre-trained by using two tasks. It hides some of the words (15%, although this is actually more complex) and learns how to find them. This allows him to acquire a general and bi-directional knowledge of the language. BERT also learns to recognize if two sentences are consecutive or not. The corpus used for this pre-training was the BooksCorpus with 800M words and a version of the English Wikipedia with 2,500M words. When it came out, BERT was be able to outperform state of the art models for a large set of NLP tasks such as GLUE (General Language Understanding Evaluation) or SQuAD (Stanford Question Answer Dataset).

In the following we use transfer learning (with our vigilance data) on two pre-trained models based on the BERT architecture and French corpus.

4.3.1 CamemBERT

CamemBERT (Facebook and Inria) is a BERT type model, pre-trained on 138GB of French text. The difference between the two models lies in their pre-training. CamemBERT has been pre-trained on a French-speaking corpus and with different hyper-parameters discovered and tested successfully for the first time by the Facebook team.

Table 5 Emergency Person Rescue prediction scores

Method	MAE	RMSE
Mean	1.9856	2.4554
Persistence	1.3504	1.8292
Mean / hour	1.6485	2.1160

4.3.2 FlauBERT

FlauBERT [31] is another French Bert trained on a large heterogeneous French corpus and its performances compared to CamemBERT are very close. The results obtained with both models show that a French language model improves the results compared to similar BERT (multilingual) models.

5 Results and Discussion

blue First, we present the scores of the univariate time series models for the prediction of the number of interventions. The models perform constant predictions equal to the mean, predictions corresponding to the persistence model (we predict the last known values), and predictions corresponding to the mean per hour (see Figure 4). Next, we develop multivariate time series models, based on one of the most powerful learning machine tools available today: XGBoost [32]. The data used are composed of calendar features (day in the week, month, year...), quantitative meteorological variables, corresponding to the three meteorological stations closest to the Doubs, which makes it possible to measure the extent to which predictions of firefighting interventions can be made by taking only numerical data from meteorological information, and only vigilance levels from bulletins as indicators. Then, we reframe the problem as a multilabel classification task in order to analyze the impact of the tabular data previously described but with a new perspective that allows us to recognize the periods with intervention peaks and not the number of interventions. For this purpose, we compared the XGBoost classifier and the one from a technique also recognized for its speed and robustness, called Random Forest [33]. Finally, we replaced the tabular data by the meteorological texts of the bulletins and applied NLP techniques to study how valuable these data can be.

5.1 bluePrediction of interventions using basic univariate time series models

Scores corresponding to constant predictions are given in this sub-section. The results obtained are shown in the Table 5 for the Emergency Person Rescue, Table 6

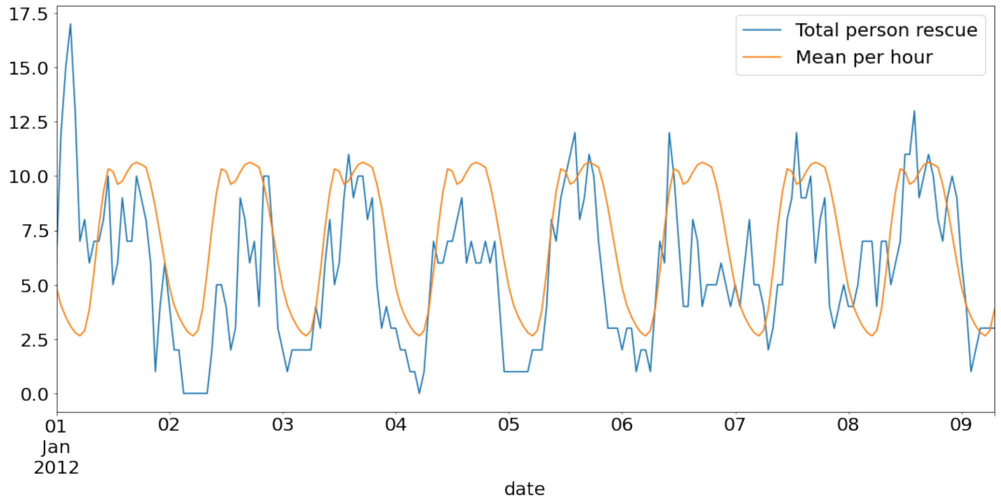


Fig. 4 Total Person Rescue interventions and its mean per hour series, early 2012

Table 6 Total Person Rescue prediction scores

Method	MAE	RMSE
Mean	3.6939	4.5267
Persistence	2.0571	2.7298
Mean / hour	2.5843	3.3901

Table 7 Heating prediction scores

Method	MAE	RMSE
Mean	0.4829	0.6240
Persistence	0.1737	0.4490
Mean / hour	0.4641	0.6127

Table 8 Storm and Flood prediction scores

Method	MAE	RMSE
Mean	0.4340	1.1742
Persistence	0.1749	0.5901
Mean / hour	0.4225	1.1699

for the Total Person Rescue, Table 7 for the Heating related interventions, and Table 8 for the Storm/Flood ones. In these tables, MAE stands for Mean Absolute Error, and RMSE is for the Root Mean Squared Error, the most usual metrics for regression. As can be seen, the average per hour does better than the average alone for personal assistance, which is understandable given the daily seasonality of human activity. As shown in Figure 4, there are fewer interventions at night than during the day because people are simply sleeping; similarly, because people eat at noon, there is a plateau at that time. This improvement is obviously not found for interventions such as heating or storm, as these are only minimally related to human activity: a storm can cause damage both day and night.

One might a priori be astonished that the simplistic persistence model does better than the average per hour, which seems more evolved. However, this is well explained by the scarcity of interventions of the heating or storm type: if, most of the time, there are 0 interventions per hour, then the probability that the slots at time t and time t+1 both have no intervention is high. Thus, replicating what happened at time t as a prediction of what will happen at time t+1 is a winning strategy when events are rare. The good success of this model in the case of rescue-type interventions is again explained by the strong daily seasonality of these interventions: between one hour and the next, the number of interventions is close, when it is very different between time t and t+12 hours. In other words, there is a strong correlation between the time series of the rescue-type and the time series shifted by one hour, as can be seen on the auto-correlation graph of the Total People Rescue, see Figure 5.

5.2 bluePrediction of interventions using multivariate time series models

The first features to be taken into consideration are obviously the calendar data, which fully condition human activity. The time of day captures the daily seasonality, with a drop in activity at night, a maximum of activity in the late morning and in the afternoon, with a plateau at noon. For various reasons, the day in the week also has its importance: weekend, day without school for children... In the same way, the day in the year makes it possible to recover particular periods such as the summer, the winter vacations, even particular days (national holiday, Christmas and New Year's Day...). These particularities are also contained, but in

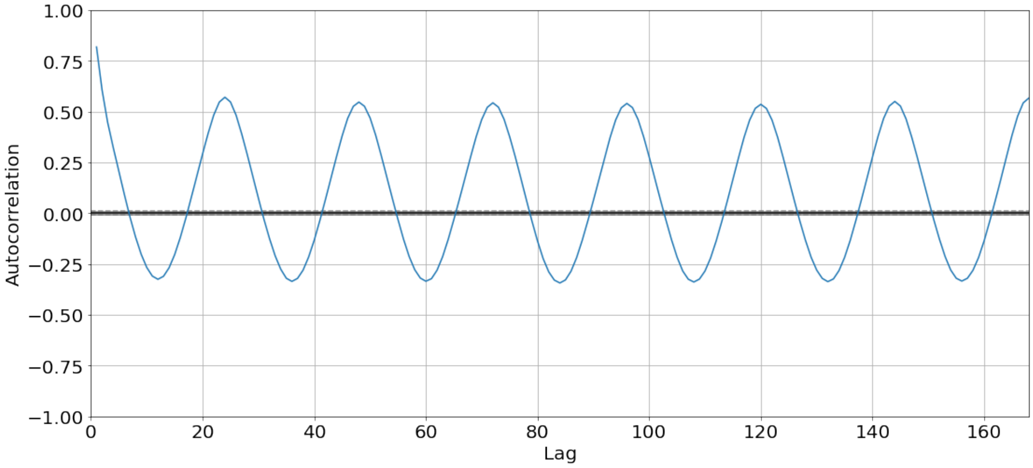


Fig. 5 Auto-correlation graph for Total Person Rescue

a less continuous way, in the month in the year. Finally, the year is important because it allows us to model the general upward trend, for the various reasons mentioned above (aging population, disengagement of the private sector ...). For accidents related to heating (chimney fire, etc.), storms and floods, or even emergency rescue, it is reasonable to think that meteorology is important, and that predictions should be improved by adding variables describing it. For example, between two national holidays, if the first one is rainy when the second one takes place under a radiant sun during a heat wave, this will probably have an impact on firefighters' exits.

Table 9 Prediction scores using XGBoost, Emergency Person Rescue case

Features	MAE	RMSE
Calendar	1.523	1.961
Weather	1.825	2.307
Vigilance	2.021	2.493
Weather + Calendar	1.586	2.049
Calendar + Vigilance	1.400	1.881
Weather + Vigilance	2.186	2.809
All	1.940	2.632

The purpose of this section is to see if this meteorological influence can be recovered without using NLP. We will therefore compare the quality of the predictions for the 4 types of intervention, with or without the temporal data (calendar), with or without the color of the weather alert (vigilance), and with or without the quantitative data from the Météo-France site (weather) such as temperature, pressure, wind, etc. To do so, we will randomly separate our data set between learning (80%) and test (20%), and we will look at the MAE and RMSE scores on the test set, once the learning is completed. This learning will be done with XGBoost (Poisson regression, max depth of 6), with 20% of the learning set used for validation, and an early stopping criterion of 10 steps.

Table 10 Prediction scores using XGBoost, Total Person Rescue case

Features	MAE	RMSE
Calendar	2.223	2.917
Weather	3.281	4.138
Vigilance	4.117	5.138
Weather + Calendar	2.375	3.118
Calendar + Vigilance	2.279	2.998
Weather + Vigilance	4.293	5.376
All	3.385	4.445

The first lesson to be learned from the Table 9 is that, in the case of emergency person rescue, calendar data is obviously what appears to be most important, but that predictions can be improved by adding the color of the weather alert bulletin. These bulletins are basically too coarse a data set to be useful in predicting these types of interventions on their own. As such, quantitative meteorological data do a little better, but are far from what is obtained with calendar data. However, the best result is obtained by mixing calendar data with vigilance bulletins, when the addition of quantitative weather variables systematically lowers the quality of the predictions. The information that they carry on their own is found in the couple of calendar and vigilance data, which come as if cleaned of the noise that the quantitative variables carry: variables such as gust over a period or mean wind speed (10 min.) have too local a scope, both spatially and temporally, whereas the bulletins have a more general scope (the information is digested). Note again that, among baselines, one does not do better there than the persistence model. An autoregressive component is fundamental to this prediction problem, and it would improve the scores of the Table 9 in an obvious way: among the interventions between t and t+1, there are the new interventions that appear after t, and those that appeared previously but are still in use (the system has a strong inertia).

Table 11 Prediction scores using XGBoost, heating case

Features	MAE	RMSE
Calendar	0.413	0.558
Weather	0.502	0.663
Vigilance	0.510	0.659
Weather + Calendar	0.475	0.636
Calendar + Vigilance	0.319	0.495
Weather + Vigilance	0.565	0.825
All	0.533	0.804

The same lessons can be learned from total interventions in a more pronounced way, see Table 10. This time, the best result is achieved by calendar data alone, and the addition of weather variables only pushes XGBoost to overfitting. These total interventions include, in addition to the emergency rescue, accidents on the public highway, and non-emergency rescue: person trapped in an elevator or locked in a balcony, wasp's nest, etc. These interventions are, by nature, much more difficult to predict. And the impact of temperature or atmospheric pressure is certainly much less on this type of intervention than, for example, knowing whether it is the middle of the day or the middle of the night. One can then naturally wonder if adding textual information on the weather could improve such results.

Table 12 Prediction scores using XGBoost, Storm/Flood case

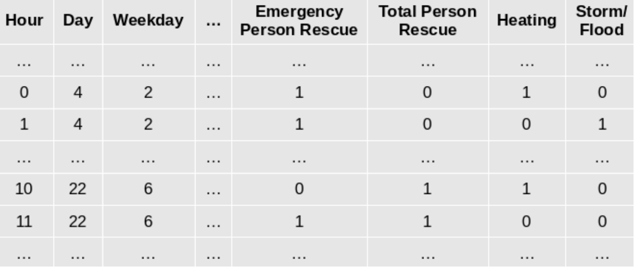
Features	MAE	RMSE
Calendar	0.325	0.734
Weather	0.386	0.867
Vigilance	1.729	4.959
Weather + Calendar	0.370	0.863
Calendar + Vigilance	0.793	2.453
Weather + Vigilance	1.212	2.565
All	1.161	2.978

In the case of interventions related to heating, this time the best score is obtained by linking the calendar data to the vigilance bulletins, as shown in Table 11. This is understandable, given the nature of the interventions (chimney fire, electrical heating appliance fire...). Here again, quantitative meteorological information does not provide the same benefit as the color of the vigilance bulletin, for the same reasons as previously mentioned. Conversely, for events such as storms and floods, if the calendar alone produces the best results, it is closely followed by the combination of calendar and meteorological data (cf. Table 12). Here, in a counter-intuitive way, the vigilance bulletins greatly reduce performance. All this can be explained by noting first of all that floods do not occur at any time of the year in the Doubs, but mainly in winter. They are very localized in time, when the bulletins of vigilance usually extend over a fairly long period. Finally, such events follow a strong fall of water, a fact that is found in weather variables.

To conclude, the case of heating-type interventions shows that by adding information about the weather, predictions can be improved. The case of storms and floods, on the other hand, shows that quantitative information (temperature, pressure...) and qualitative information (risk of storm, flooding...) do not provide enough information on the weather: the latter has an obvious impact on the interventions, but these variables do not improve the predictions. And the Emergency and Total People Rescue confirm that these variables describe the weather forecast too crudely, which explains why we are interested in the textual content of the vigilance bulletins.

5.3 bluePrediction of intervention peaks using multilabel classification models based on decision trees and tabular data

blue In this section, the problem is restated as a multilabel classification, where the labels are the 4 categories Emergency Person Rescue, Total Person Rescue, Heating, and Storm/Flood, and the possible classes are 0 and 1 calculated as described in 3.2. This is in order to identify the peak periods of intervention with respect to each category. The quantitative and qualitative variables described in the preceding section are the inputs for the models developed in this section, where each sample of the dataset represents one hour, an illustration is showed in Figure 6. In addition, the models created are the results of various combinations of the variables with the objective of observing the influence of them but with a classification approach. These models will be the baselines to be overcome by the NLP techniques in the following section.



Hour	Day	Weekday	...	Emergency Person Rescue	Total Person Rescue	Heating	Storm/ Flood
...	...	...	...	...	...	...	...
0	4	2	...	1	0	1	0
1	4	2	...	1	0	0	1
...	...	...	...	...	...	...	...
10	22	6	...	0	1	1	0
11	22	6	...	1	1	0	0
...	...	...	...	...	...	...	...

Fig. 6 Illustration example of the tabular data for multilabel classification, considering calendar variables

Before applying NLP techniques, we also sought to develop and compare models with simpler techniques, which do not require a greater consumption of resources, such as those based on decision trees, but which have demonstrated their remarkable effectiveness in the literature with tabular data. We chose XGBoost as a representative of boosting algorithms, which seek to reduce model variance, and Random Forest, as a representative of bagging algorithms, which seek to reduce bias.

Since decision trees are robust to work with categorical and continuous values, we kept the original values of our variables. Furthermore, similar to the previous section, we randomly split the data into 80% for training and 20% for testing. Since XGBoost and Random Forest do not natively support multitarget classification, we used *MultiOutputClassifier*, from the Scikit-Learn library, to fit one classifier per target. To select the best model, we used Bayesian optimization via the HyperOpt library. In total 100 iterations were performed for each combination of variable type and for each technique, and to guide the search for the best model we set as loss function the metric *Micro F1-score*, hereafter called *F1-score* for short, generally used to assess the quality of multilabel binary problems, where the score closer to 1 means better *Micro-Precision* and *Micro-Recall* (we will abbreviate both to *Precision* and *Recall*, respectively) and closer to 0 means poor model performance. Other metrics such as *Accuracy* and *Balanced Accuracy* are also presented to analyze our resulting models.

Table 13 shows the results of the best models obtained for each data input combination and Table 16 describes their hyperparameters and the search space used. As we can see in Table 13, calendar data together with vigilance alert levels improve the performance of the models. What is more, the best model, obtained by Random Forest, used only calendar data reaching an F1-score of 0.81. Also, we note that when the inputs are weather, vigilance indicators, and weather plus vigilance indicators, the models show an F1-score below 0.64, a poor performance, when in fact these variables should present a greater contribution in the recognition of intervention peaks.

Therefore, these are the basic results that we need to outperform to be efficient. An NLP tool must have F1-score at least greater than those values. So, we have to see whether information derived from the text of the bulletins, make it possible to better predict interventions than simple calendar data, vigilance levels, or quantitative information such as temperature or wind speed. For this reason, we are interested in comparing different NLP models.

5.4 bluePrediction of intervention peaks using multilabel classification models based on NLP techniques and text from meteorological bulletins

blue In this section, we seek to discover if we have better learning for the 4 outputs when we process the texts of the bulletins.

For this task, the dataset used considers the bulletins of the three neighboring departments of Doubs and the bulletins of Doubs. They were structured as shown in Table 2 and the task is maintained as a multilabel classification for the recognition of the intervention peak periods.

The initial dataset was split into 80% for the learning phase and 20% for the testing phase. Experimentation was then performed by varying different hyperparameters for the different models. blueResults presented in Table 14 correspond to the best results obtained for each technique applied. We calculated the same metrics presented in the previous section 5.3 to identify possible improvements. On the one hand, as described in sections 4.1 and 4.2, for LSTM and CNN, we applied a text preprocessing and used the library Hyperopt to search for the best configuration for our models, where the number of iterations was 100, and the guiding metric was the F1-score. The best setting for LSTM was the architecture n$^{\text{o}}$1, with a learning rate = 6e-4, a number of epochs = 200, and the batch size = 59. The best configuration for CNN was the architecture n$^{\text{o}}$3 with a learning rate = 9e-3, a number of epochs = 105, and the batch size = 95. On the other hand, transformers were used in its base models: CamemBERT (110 Million parameters) and FlauBERT (138 Million parameters) with the following hyperparameters. For CamemBERT: learning rate = 1e-5, number of epochs = 75, and batch size = 48. For FlauBERT: learning rate = 1e-5, number of epochs = 150, and batch size = 128. The search of the best hyperparameters was performed by using a random search with an early stopping strategy of 15 iterations. We used Huggingface implementations for both CamemBERT and FlauBERT models. For more details on the description of the search space and the best configuration, see Table 17.

blue From Table 14, we see that LSTM and CNN results are quite similar. However, the two French transformers models, CamemBERT and FlauBERT, outperform both traditional techniques. These set of experiments confirm the recent literature results [34] on text classification problems and the superiority of Transformers models. Note that, Accuracy and Balanced Accuracy are quite different due to the imbalance of the dataset as mentioned in Section 3.2. Nevertheless, when looking at the F1-score, all the models in this section

Table 13 Prediction results of the multilabel models based on XGBoost and Random Forest techniques

Technique	Input	F1-score	Accuracy	Balanced Accuracy	Precision	Recall
XGBoost	Calendar	0.80	0.48	0.82	0.80	0.80
	Weather	0.63	0.21	0.64	0.65	0.60
	Vigilance	0.62	0.23	0.62	0.61	0.62
	Weather + Calendar	0.71	0.30	0.71	0.73	0.69
	Calendar + Vigilance	0.81	0.48	0.82	0.81	0.80
	Weather + Vigilance	0.64	0.24	0.65	0.65	0.64
	All	0.71	0.35	0.71	0.74	0.68
Random Forest	Calendar	**0.81**	**0.51**	**0.83**	**0.82**	**0.81**
	Weather	0.61	0.24	0.61	0.63	0.59
	Vigilance	0.64	0.18	0.58	0.54	0.77
	Weather + Calendar	0.73	0.35	0.72	0.74	0.71
	Calendar + Vigilance	0.81	0.50	0.83	0.81	0.81
	Weather + Vigilance	0.62	0.24	0.61	0.63	0.61
	All	0.72	0.33	0.72	0.74	0.71

Table 14 Prediction results of the multilabel models based on NLP techniques

Technique	Input	F1-score	Accuracy	Balanced Accuracy	Precision	Recall
CNN	Bulletin Text	0.84	0.56	0.78	0.83	0.85
LSTM		0.85	0.56	0.80	0.84	0.86
FlauBERT		0.87	0.59	0.82	0.86	0.89
CamemBERT		**0.89**	**0.65**	**0.84**	**0.87**	**0.90**

Table 15 Accuracy results for each type of intervention, considering the models generated with the weather bulletins.

Model	Emergency Person Rescue	Total Person Rescue	Heating	Storm/Flood
CNN	0.76	0.82	0.84	0.82
LSTM	0.76	0.83	0.85	0.83
FlauBERT	0.79	0.87	0.88	0.85
CamemBERT	**0.80**	**0.86**	**0.92**	**0.86**

remarkably outperform the results obtained in the previous section with decision trees (Table 13). What is more, the best model obtained with CamemBERT overcome by far the best model obtained with Random Forest by 8%, 14%, 1%, 5%, and 9%, when comparing F1-score, Accuracy, Balanced Accuracy, Precision, and Recall, respectively.

When we examine the accuracy by independent category, in Table 15, we prove that extracting features from the texts enhance the recognition of intervention peaks due to rare events, since the best NLP model reached accuracies of 80%, 86%, 92%, and 86% for Emergency Rescue People, Total Rescue People, Heating, and Storm/Flood, respectively. Moreover, the last 2 categories Heating and Storm/Flood, that represent interventions generated by rare events and were complicated to recognize by the approaches analyzed in the previous sections, are the ones that demonstrated high accuracy with the NLP techniques and bulletin texts, without degrading the accuracy of the 2 rescue-type categories.

As mentioned in section 2, generally, the models developed for predicting the interventions of fire departments or EMS use tabular data such as temporal information, quantitative meteorological variables, traffic indicators, etc [3,5,6,11]. These are very useful for recognizing incidents related to human activity, since it is possible to identify seasonality and trend over time (people are more active during the day than at night, there are more drownings in the pool during the summer than the winter, as the population increases the incidents also increase, etc.), and because the major operational burden of these organizations is the interventions rescue-type. However, there are interventions that are difficult to detect, since they occur only a few times a year. These are produced by rare events such as natural phenomena (storms, floods, forest fires, etc.), and although their occurrence is minimal over the years, the workload they produce in a small period of time can be 28 times more than normal in some cases. For example, in 2016, the average number of interventions assisted by SDIS 25 per hour was 3.34, however, there was an hour in which 84 interventions occurred, due to a storm. This caused flooding in the region, human and material damage, and breakdowns in the service of SDIS 25 [3]. For this reason, we need an intelligent system that could help to predict the peak periods of intervention generated by rare events. Thus, the present work developed models based on NLP techniques and meteorological bulletins from public sources to recognize periods with a heavy operational load. The results obtained are significant for practical purposes. Initially, the model could be deployed in production as a small stand-alone application. Or, the predictions (the binary indicators) could be included in a larger set of tabular data that would be the input for a predictive model of the number of interventions for a certain time period and location. In this way, with an initial application or with a more robust system, the fire department could reorganize its personnel and armament to cope with these periods of high demand, reduce breakdowns in service due to lack of resources, and save more lives.

6 Conclusion

blue The present paper demonstrates the effectiveness of NLP techniques for the recognition of rare events that will cause an increase above the average in certain periods of firefighter interventions. This is done by processing the texts contained in the weather bulletins using the traditional techniques LSTM and CNN, and transformers CamemBERT and FlauBERT. The results of the NLP models and bulletin texts exceed those of the baselines with Decision Trees (XgBoost and Random Forest) and tabular data by 8% and 14% when comparing the best F1-score and Accuracy, respectively. The advantage of using these texts is also reflected when assessing the accuracy of the 2 categories with interventions related to rare events, achieving 92% for Heating and 86% for Storm/Flood with the best CamemBERT model developed.

In this way, fire departments and EMS, in general, would be able to identify peak periods of interventions and optimize their response by establishing better strategies to prepare their armament for natural disasters (storms, floods, etc.) and keep the population better protected and safe.

As future work, we propose to add meteorological bulletins from other departments of the country, which would allow us to better track a possible extreme event and its consequences on the firefighters' workload. Furthermore, we will develop and compare other modeling and text preprocessing techniques with the texts in French and English. Finally, we aim to integrate the results of this classification approach into a larger regression model that predicts the number of interventions for a certain time period (hourly, daily and monthly) and by locality (principal cities and mountain cities).

Acknowledgments and funding

This work was funded by the Service Départemental d'Incendie et de Secours du Doubs (SDIS 25), with the support of the French Ministry of Higher Education and Research (managed by the National Association of Research and Technology (ANRT) for the CIFRE thesis (N 2019/0372), and by the EIPHI-BFC Graduate School (contract "ANR-17-EURE-0002"). Computations have been performed on the supercomputer facilities of "Mésocentre de Calcul de Franche-Comté".

Conflicts of interest/Competing interests

Authors certify that there is no conflicts of interests or competing in interests in these works.

References

1. S. Cerna, C. Guyeux, G. Royer, C. Chevallier, and G. Plumerel, "Predicting fire brigades operational breakdowns: A real case study," *Mathematics*, vol. 8, no. 8, p. 1383, Aug. 2020.
2. H. H. Arcolezi, J.-F. Couchot, S. Cerna, C. Guyeux, G. Royer, B. A. Bouna, and X. Xiao, "Forecasting the number of firefighter interventions per region with local-differential-privacy-based data," *Computers & Security*, vol. 96, p. 101888, Sep. 2020.
3. S. Cerna, C. Guyeux, H. H. Arcolezi, R. Couturier, and G. Royer, "A comparison of LSTM and XGBoost for predicting firemen interventions," in *Trends and Innovations in Information Systems and Technologies. WorldCIST 2020.* Springer International Publishing, 2020, vol. 1160, pp. 424–434.
4. K. Pirklbauer and R. D. Findling, "Predicting the category of fire department operations," in *Proceedings of the 21st International Conference on Information Integration and Web-based Applications & Services.* ACM, Dec. 2019.
5. T. F. Morello, R. M. Ramos, L. O. Anderson, N. Owen, T. M. Rosan, and L. Steil, "Predicting fires for policy making: Improving accuracy of fire brigade allocation in the brazilian amazon," *Ecological Economics*, vol. 169, p. 106501, Mar. 2020.
6. S. H. Sankar, K. Jayadev, B. Suraj, and P. Aparna, "A comprehensive solution to road traffic accident detection and ambulance management," in *2016 International Conference on Advances in Electrical, Electronic and Systems Engineering (ICAEES).* IEEE, Nov. 2016. [Online]. Available: https://doi.org/10.1109/icaees.2016.7888006
7. A. Y. Chen, T.-Y. Lu, M. H.-M. Ma, and W.-Z. Sun, "Demand forecast using data analytics for the preallocation of ambulances," *IEEE Journal of Biomedical and Health Informatics*, vol. 20, no. 4, pp. 1178–1187, Jul. 2016.

8. I. F. de la Mota, E. S. Perez, and A. V. Garduno, “Optimization and simulation of an ambulance location problem,” in *2017 Winter Simulation Conference (WSC)*. IEEE, Dec. 2017. [Online]. Available: https://doi.org/10.1109/wsc.2017.8248188
9. A. Carvalho, M. Captivo, and I. Marques, “Integrating the ambulance dispatching and relocation problems to maximize system’s preparedness,” *European Journal of Operational Research*, vol. 283, no. 3, pp. 1064–1080, Jun. 2020.
10. R. I. Rasel, N. Sultana, G. A. Islam, M. Islam, and P. Meesad, “Spatio-temporal seismic data analysis for predicting earthquake: Bangladesh perspective,” in *2019 Research, Invention, and Innovation Congress (RI2C)*. IEEE, Dec. 2019.
11. X. S. Lu, M. Zhou, and L. Qi, “Analyzing temporal-spatial evolution of rare events by using social media data,” in *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, Oct. 2017.
12. B. L. Cook, A. M. Progovac, P. Chen, B. Mullin, S. Hou, and E. Baca-Garcia, “Novel use of natural language processing (NLP) to predict suicidal ideation and psychiatric symptoms in a text-based mental health intervention in madrid,” *Computational and Mathematical Methods in Medicine*, vol. 2016, pp. 1–8, 2016.
13. M. R. Hasan, M. Maliha, and M. Arifuzzaman, “Sentiment analysis with NLP on twitter data,” in *2019 International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2)*. IEEE, Jul. 2019.
14. Y. Ding, B. Li, Y. Zhao, and C. Cheng, “Scoring tourist attractions based on sentiment lexicon,” in *2017 IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*. IEEE, Mar. 2017.
15. A. K. M. Rasu, *Hands-On Python Natural Language Processing: Explore tools and techniques to analyze and process text with a view to building real-world NLP applications*. Packt Publishing Ltd, 2020.
16. S. Preum, S. Shu, M. Hotaki, R. Williams, J. Stankovic, and H. Alemzadeh, “CognitiveEMS,” *ACM SIGBED Review*, vol. 16, no. 2, pp. 51–60, Aug. 2019.
17. I. Temnikova, C. Castillo, and S. Vieweg, “Emterms 1.0: A terminological resource for crisis tweets,” in *ISCRAM*, 2015.
18. D. Kozlowski, E. Lannelongue, F. Saudemont, F. Benamara, A. Mari, V. Moriceau, and A. Boumadane, “A three-level classification of french tweets in ecological crises,” *Information Processing & Management*, vol. 57, no. 5, p. 102284, Sep. 2020.
19. A. Karami, V. Shah, R. Vaezi, and A. Bansal, “Twitter speaks: A case of national disaster situational awareness,” *Journal of Information Science*, vol. 46, no. 3, pp. 313–324, Mar. 2019.
20. W. J. Corvey, S. Vieweg, T. Rood, and M. Palmer, “Twitter in mass emergency: what nlp techniques can contribute,” in *HLT-NAACL 2010*, 2010.
21. E. Fersini, E. Messina, and F. A. Pozzi, “Earthquake management: a decision support system based on natural language processing,” *Journal of Ambient Intelligence and Humanized Computing*, vol. 8, no. 1, pp. 37–45, Apr. 2016.
22. Y. Sermet and I. Demir, “An intelligent system on knowledge generation and communication about flooding,” *Environmental Modelling & Software*, vol. 108, pp. 51–60, Oct. 2018.
23. “Météo-France public data,” https://donneespubliques.meteofrance.fr/?fond=produit&id_produit=90&id_rubrique=32, accessed: 2021-02-01.
24. “Vigilance cards and bulletins archive, Météo-France,” http://vigilance-public.meteo.fr/index.php, accessed: 2021-02-01.
25. S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, nov 1997.
26. I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.
27. D. T. Nguyen, K. A. A. Mannai, S. Joty, H. Sajjad, M. Imran, and P. Mitra, “Rapid classification of crisis-related data on social networks using convolutional neural networks,” 2016.
28. J. Pennington, R. Socher, and C. D. Manning, “Glove: Global vectors for word representation,” in *Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543. [Online]. Available: http://www.aclweb.org/anthology/D14-1162
29. T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” 2013.
30. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” 2019.
31. H. Le, L. Vial, J. Frej, V. Segonne, M. Coavoux, B. Lecouteux, A. Allauzen, B. Crabbé, L. Besacier, and D. Schwab, “Flaubert: Unsupervised language model pre-training for french,” 2020.
32. T. Chen and C. Guestrin, “Xgboost,” *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug 2016.
33. L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
34. S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, “Deep learning based text classification: A comprehensive review,” 2021.

A Settings used during modeling

Table 16 Hyperparameters search space and the best configuration for XGBoost and Random Forest multilabel models

Method	Search Space	Best configuration						
		Calendar	Weather	Vigilance	Weather + Calendar	Calendar + Vigilance	Weather + Vigilance	All
XGBoost	n_estimators: [50-200]	195	159	193	132	132	86	112
	learning_rate: [0.001-0.8]	0.49	0.14	0.39	0.17	0.40	0.60	0.69
	max_depth: [1-100]	100	6	2	3	20	1	4
	colsample_bytree: [0.2-1]	0.99	0.5	0.56	0.42	0.87	0.22	0.46
	objective: multi:softmax	multi:softmax						
	eval_metric: mlogloss	mlogloss						
Random Forest	n_estimators: [50-500]	52	328	61	152	411	337	87
	max_features: [0.2-1]	0.38	0.74	0.88	0.46	0.53	0.94	0.22
	max_depth: [1-10]	100	1	10	5	20	1	5
	class_weight: [balanced, balanced_subsample]	balanced_subsample	balanced	balanced	balanced	balanced	balanced	balanced

Table 17 Hyperparameters search space and the best configuration for NLP multilabel models

Method	Search Space	Best configuration
CNN	type of architecture: [1,2,3]	3
	learning rate: [0.00001-0.01]	0.009
	batch size: [40-150]	95
	epochs: 500	105
	early stopping: 15	15
	restore best weights: True	True
LSTM	type of architecture: [1,2,3]	1
	learning rate: [0.00001-0.009]	0.0006
	batch size: [40-150]	59
	epochs: 200	200
	early stopping: 20	20
	restore best weights: True	True
FlauBERT	type of architecture: flaubert-base-cased	flaubert-base-cased
	learning rate: [0.0001, 0.00001]	0.00001
	batch size: [16-256]	128
	epochs: [10-200]	150
	early stopping: 15	15
	restore best weights: True	True
CamemBERT	type of architecture: camembert-base	camembert-base
	learning rate: [0.0001, 0.00001]	0.00001
	batch size: [16-256]	48
	epochs: [10-200]	75
	early stopping: 15	15
	restore best weights: True	True