ARTICLE

Extracting Natech Reports from Large Databases: Development of a Semi-Intelligent Natech Identification Framework

Xiaolong Luo¹ · Ana Maria Cruz² · Dimitrios Tzioutzios¹

Accepted: 10 October 2020 / Published online: 6 November 2020
© The Author(s) 2020

Abstract Natural hazard-triggered technological accidents (Natechs) refer to accidents involving releases of hazardous materials (hazmat) triggered by natural hazards. Huge economic losses, as well as human health and environmental problems are caused by Natechs. In this regard, learning from previous Natechs is critical for risk management. However, due to data scarcity and high uncertainty concerning such hazards, it becomes a serious challenge for risk managers to detect Natechs from large databases, such as the National Response Center (NRC) database. As the largest database of hazmat release incidents, the NRC database receives hazmat release reports from citizens in the United States. However, callers often have incomplete details about the incidents they are reporting. This results in many records having incomplete information. Consequently, it is quite difficult to identify and extract Natechs accurately and efficiently. In this study, we introduce machine learning theory into the Natech retrieving research, and a Semi-Intelligent Natech Identification Framework (SINIF) is proposed in order to solve the problem. We tested the suitability of two supervised machine learning algorithms, namely the Long Short-Term Memory (LSTM) and the Convolutional Neural Network (CNN), and selected the former for the development of the SINIF. According to the results, the SINIF is efficient (a total number of 826,078 records were analyzed) and accurate (the accuracy is over 0.90), while 32,841

Natech reports between 1990 and 2017 were extracted from the NRC database. Furthermore, the majority of those Natech reports (97.85%) were related to meteorological phenomena, with hurricanes (24.41%), heavy rains (19.27%), and storms (18.29%) as the main causes of these reported Natechs. Overall, this study suggests that risk managers can benefit immensely from SINIF in analyzing Natech data from large databases efficiently.

Keywords Data extraction method · Machine learning · Natechs · Natural hazards · NRC database

1 Introduction

Natural hazard-triggered technological accidents are known as Natechs (Showalter and Myers 1992; Cruz et al. 2006; Cruz and Okada 2008). In this study, we are interested in the Natechs that involve facility/equipment damage with the subsequent release of hazardous materials (hazmat). Natechs can cause huge economic losses (Girgin and Krausmann 2016; Krausmann and Salzano 2017) as well as long-term effects on human health and the environment (Krausmann and Cruz 2013). Natechs are generally considered as hazmat release accidents that occur from the impact of a natural hazard on vulnerable industrial infrastructures (for example, storage tanks, fixed facilities, oil drill platforms). As a result, Natechs are more complex, and can have more severe consequences than the triggering natural hazard alone. Therefore, Natechs bring a huge challenge to risk managers.

In this regard, learning from previous Natechs is crucial for risk analysis, assessment, and reduction. Over the past decades, researchers have studied Natech risk assessment methodologies and the vulnerability of storage tanks

✉ Ana Maria Cruz
cruznanajo.anamaria.2u@kyoto-u.ac.jp

¹ Department of Urban Management, Graduate School of Engineering, Kyoto University, Kyoto 615-8540, Japan

² Disaster Prevention Research Institute, Kyoto University, Kyoto 611-0011, Japan

(Antonioni et al. 2007; Campedel et al. 2008; Antonioni et al. 2009, 2015; Khakzad and Van Gelder 2017, 2018; Khakzad et al. 2018). Meanwhile, many studies attempted to understand the incidence and characteristics of these events by historically analyzing Natechs (Krausmann and Mushtaq 2008; Cruz and Krausmann 2009; Cozzani et al. 2010; Krausmann and Cruz 2013; Girgin and Krausmann 2014, 2016; Kumasaki et al. 2017; Shah et al. 2018). Some of the above studies have used several databases that contain records of chemical release accidents (including Natechs) to support research concerning Natech-related issues, such as the National Response Center (NRC) database of the United States (United States Coast Guard 2017), the eNatech database of the Joint Research Centre of the European Commission (European Commission 2019a), the Failure and Accidents Technical information System (FACTS) (TNO Industrial and External Safety 2019), the Research and Information on Accidents database (ARIA) (Bureau for Analysis of Industrial Risk and Pollution 2019), and the major accidents reporting systems (eMARS) database (European Commission 2019b), among others. Among these databases, the eNatech is the sole database specifically dedicated to Natechs. However, due to the fact that eNatech contains only few records (just over 60 records), and the collection of accidents is not systematic, this database is not very useful for analyzing Natech incidence over a territory and over time. Therefore, it is necessary to use other available and larger databases for this purpose. However, with larger databases come greater challenges.

Previous studies have retrieved Natechs from the “*Description*” field in the available databases. To do this, the researchers have to filter appropriate data according to the research purpose. For example, in a study by Krausmann et al. (2011), the authors first extracted records regarding certain types of equipment (for example, storage tanks, pipelines) and their hazmat release cause, and then reviewed the description of the incident field manually for each record to obtain more details regarding the incident causes. In their analysis the authors reviewed several databases. The total number of Natech events identified was just over 1000, which must have taken some time to review by hand. The study of Krausmann et al. (2011) provided valuable insight concerning the extraction of Natech cases from databases, but also highlighted that this task is heavily time consuming, particularly if a larger set of incidents needs to be reviewed. This is the case when we attempt to retrieve a complete set of Natechs from a million-volume level database, such as the complete NRC database.

The NRC database is operated by the United States Coast Guard, receiving and cataloguing all citizens’ reports for potential chemical release incidents. Since its

establishment in 1974 (Clow 1980), the NRC database has evolved over the years, receiving approximately 25,000 to 35,000 hazmat release reports per year. The NRC database contains records from 1990 onwards, and it is open and available to the public. Furthermore, these reports, which can be downloaded from the NRC website, are released as separate files that contain all hazmat release reports recorded per year. During the period of 1990 to 2017, the total number of reports exceeded 820,000. According to Girgin and Krausmann (2014), there are over 110,000 records from 1983 to 1989. Compared with other available databases, the NRC database contains by far the largest number of reports about chemical release incidents. Due to the large amount of data, it is almost impossible to retrieve Natechs by checking the description one by one. Moreover, the NRC database is increasing by almost 29,000 records on average per year. This is near to the total size, for example, of the FACTS database, half of the total number of records in the ARIA database, and almost 29-fold the total number of records in the eMARS database. Also the number of potential chemical release incident and accident reports in all available databases per year is increasing. As these databases expand, it will be useful to find a fast and efficient method to retrieve Natechs from large databases for accident analysis.

Improving Natech identification in the NRC database is challenging also because the NRC database has many inherent problems (Girgin and Krausmann 2014). First, there are many different types of reports catalogued in the NRC database, such as: (1) reports of planned hazmat release incidents; (2) accident reports without hazmat release; (3) accident reports with hazmat release unrelated to natural phenomena; and (4) accident reports with hazmat release related to natural phenomena (Natech events). Moreover, there may be several reports on the same incident as well as incomplete or limited details about some incidents altogether. This is in part due to the fact that any citizen can report hazmat release cases to the NRC call center. This results in many records having incomplete information. Due to the lack of knowledge and information, these redundant, complex, and confusing reports bring a huge challenge on Natech identification. In addition, changes to reporting criteria and changes to the data collection and entry forms have resulted in differences in the database itself for different periods (Cruz and Okada 2008; Krausmann and Salzano 2017). For example, since 2003, the NRC database started to include “hurricane” as one of the incident causes in the field of “*Incident cause*.”

Another problem concerns the uncertainty regarding the natural hazard cause, as the general option “natural phenomena” can also be indicated. Thus, we can deduct that one Natech was caused by a natural phenomenon, but not which particular type of natural hazard. In order to identify

these accidents, we would need to review each and every accident description, which, if done manually, would be a monumental task.

One way to identify Natechs is keyword search. A keyword extraction method has been employed to extract Natechs from the NRC database in the study by Sengul et al. (2012). Nonetheless, due to the problems introduced by language expression, it is also a time-consuming work to extract Natechs from the NRC database by checking the description of each incident. For example, using “snow” as a keyword to extract snow-triggered accidents (or “incidents” as reported in the NRC database) from the NRC database, we will get some results as shown in Table 1. All the records in Table 1 were identified using this keyword. However, none of these records were actually caused by snow. One option to avoid having this type of problem is to check every record manually, although this method is practically infeasible for large databases. Consequently, an accurate and efficient automated method of retrieving Natechs from the NRC database is needed.

For the purpose of extracting Natechs and identifying related natural phenomena from the NRC database, we introduce machine learning theory into this study. A new Natech retrieval framework, a Semi-Intelligent Natech Identification Framework (SINIF), is proposed based on a keyword extraction method and machine learning theory. A total number of 826,078 chemical release reports stored in the NRC database from 1990 to 2017 were analyzed using the proposed SINIF. In this particular study we focused on all the incidents reports involving chemical releases from fixed industrial facilities, storage tanks, processing equipment, on- and offshore platforms, and pipelines. Furthermore, we include in the analysis reports involving releases from mobile sources, such as tanker trucks, vessels, and trains.

In the following section we provide a brief introduction to machine learning, and why we selected it and how it is adapted for the purpose of this study. We then explain in detail the design and architecture of the SINIF. In Sect. 4, we introduce the methodology including data collection,

processing, and analysis workflow. The results and discussion are elaborated in Sect. 5, while the key conclusions of this study are presented in Sect. 6.

2 What is Machine Learning?

Machine learning refers to a series of algorithms and statistical models that help researchers understand complex data, especially big data. They are typically developed and implemented on computer systems to perform specific tasks effectively based on pattern recognition and statistical inference (Bishop 2006). Notable applications of machine learning include image classification (Lu and Weng 2007; Sudharshan et al. 2019), regression (Huang et al. 2011; Bashar and Mahmud 2019), and face detection and recognition (Campadelli et al. 2004; Ranjan et al. 2019). Furthermore, another application area of machine learning is Natural Language Processing (NLP). Natural Language Processing attempts to make computers understand the natural language of human beings in order to tackle and complete useful tasks (Chowdhury 2003), such as document classification (Manevitz and Yousef 2001; Rubin et al. 2012), email spam filtering (Blanzieri and Bryl 2008; Diale et al. 2019), and so forth. Machine learning, in particular, has a remarkable performance record on solving problems of text multi-classification (Le and Mikolov 2014; Sboev et al. 2016; Soththisopha and Vateekul 2018). In brief, text multi-classification methods attempt to classify series of long text or sentences into more than two categories depending on the specific purpose, such as content recognition and emotion recognition.

According to Murphy (2012), machine learning is generally divided into two principal types: supervised learning and unsupervised learning. Both approaches dictate that the models have to be initially trained in order to make the computer system “learn” or “understand” the characteristics of the input data so that these machine learning models can be used to evaluate the target dataset. However, the difference between supervised and unsupervised

Table 1 Example records in the National Response Center (NRC) database

SEQNOS ^a	Description of incident
199671	Unknown white substance on roadway/looks like snow/foul odor maybe ammonia smell
293589	Caller states that there is a foamy substance like snowflakes in air/tan in color/caller has breathing concerns
757912	Caller states there are two companies near her home one of which is somehow releasing materials that are falling from the sky that looks like glitter and is falling like snow
807281	Caller stated that they spotted a sheen that looks like snow on the water
1198261	Caller received a report of Styrofoam falling like snow from a high rise building onto a pier and into water from a unknown source

^aThe NRC database creates a key field named “SEQNOS” as incident ID to distinguish the reports with each other

learning is that the former asks whether the training data contains any predetermined labels, while the latter has no such requirement. Therefore, supervised learning is better suited to perform tasks characterized by an expected and clearly defined output, such as classification or regression problems. On the other hand, unsupervised learning is more suitable for solving problems that do not have any a priori output, such as knowledge discovery. By conceptualizing our research aim in the context of machine learning, in essence we attempt to identify the NRC incident reports using certain labels that include either “NotNatech” or the type of the natural hazard that may have caused the chemical release. These labels are then used to evaluate whether the incident fits the description of a Natech or not and what is the associated triggering natural phenomenon. Considering the above characteristics of the two learning types, it would appear that we can employ a supervised deep learning method to achieve this and retrieve Natechs from the NRC database.

The incidents reports recorded in the NRC database do have a field named “*Description of incident*” to record the incident description from the reporter. The description is organized as a long text that includes information about the cause of the incident, the weather conditions when the incident happened, the affected equipment, whether this report was a duplicate report of the previous record or not, and when the incident happened and so forth. Based on this, it is logical to assume that the records, which can be potentially classified as Natech-related incidents, will probably have similar descriptions. For example, if there are several incident reports that can be potentially classified as hurricane-triggered Natech accidents, then certain defined keywords, such as “hurricane/typhoon,” “heavy rain,” “high speed,” “strong wind,” or the name of hurricane/typhoon, should appear in their respective descriptions. This concept explains the rationale why the keyword extraction method can be used to retrieve Natech-related records from the NRC database. However, such method requires the selected keywords to appear in the description of the analyzed incident report. If the description does not contain the selected keywords, then the analyzed incident report is omitted from being identified as a Natech, unless otherwise verified (such as in the studies of Sengul et al. (2012) or Girgin and Krausmann (2014), in which the authors confirmed whether the report was related to a Natech accident or not based on an analysis of multi-source incident databases). Conversely, when applying a deep learning method to analyze the description of the incidents, each incident description will be transformed into a word bag, which is a multi-set of containing words. These transformed word bags will be analyzed by the selected algorithms, and the content similarity will be determined. This analysis allows to identify whether the descriptions

belong to one class or another, and consequently represents the final step in the Natech identification and classification stage.

Following this rationale, the process of extracting Natechs from the NRC database can be conceptualized in the following two steps. First, classifying the incident description into different categories depending on the triggering cause; secondly, filtering out and extracting the records of which the triggering accident cause is related to natural phenomena. Thus, the research purpose of this study, namely extracting Natechs from the NRC database, can be considered as a multi-classification problem. As mentioned above, machine learning algorithms based on supervised learning approaches are more suitable for solving this kind of problem. In order to decide on the appropriate machine learning algorithm for this case, we selected two commonly employed supervised learning algorithms, the Long Short-Term Memory (LSTM) and the Convolutional Neural Network (CNN), to develop the SINIF.

The LSTM was first proposed in 1997 (Hochreiter and Schmidhuber 1997), and is a kind of recurrent neural network (RNN). It was originally developed to address the challenges previous RNN faced with capturing long-term time correlations in the input data. By the time of this study, the LSTM has been widely employed in the areas of speech recognition (Fernández et al. 2007), handwriting recognition (Graves and Schmidhuber 2009), and text classification (Shih et al. 2018).

The CNN was first proposed by Fukushima (1980), who labeled it Neocognitron. A CNN network always has one or more convolutional layers. This method is commonly used in visual imagery. With the advancement of machine learning over the years, the application scope of the CNN has become much broader, coping competently with an array of complex tasks ranging from image recognition (Cireşan et al. 2012) to video analysis and human action recognition (Ji et al. 2013) and text classification (Shin et al. 2018).

3 Design of the Semi-Intelligent Natech Identification Framework (SINIF)

In order to extract Natechs by using machine learning, we designed the SINIF (see Fig. 1). The main idea of the SINIF is that the input data (incident description) will be analyzed and classified by both the keyword extraction method and the network implemented by a machine learning algorithm, separately. In more detail, every record of input data will be classified into various categories according to the triggering cause determined by both the keyword extraction method and the network, separately. In

Fig. 1 Structure of the Semi-Intelligent Natech Identification Framework (SINIF)

the final step, the category classified by the keyword extraction method will be compared with the category classified by the network for each and every record. If the two automatically generated results do not match for any of the analyzed records, then the specific record will be manually checked and identified by the researchers. Following the above idea, five procedures are designed in the SINIF and described as follows:

- (1) *Keyword extraction analysis* The input data will be analyzed by the keyword extraction method in this step. According to the keyword identified by researchers, the input data will be classified into different categories.
- (2) *Network training* The network will be built based on the implementation of the machine learning algorithm selected by researchers in this step. The network will be trained by a set of sample data that was generated from the initial input dataset.
- (3) *Network analysis* The trained network will be used to analyze input data.
- (4) *Data comparison* A scan to check if the cause identified by the keyword extraction method matches the cause identified by the network will be conducted in this step. If the two automatically identified causes do not match with each other, researchers need to inspect the accident description and identify the triggering cause based on their own knowledge and understanding.

- (5) *Output results* The Natech reports will be separated from other hazmat release incidents reports on the basis of whether the triggering cause is related to a natural phenomenon, and if so, they will be further categorized according to the type of natural hazard.

The network itself can be based on any of the machine learning algorithms available. As mentioned in Sect. 2, we used the LSTM and CNN to develop the SINIF. Researchers can choose a suitable machine learning algorithm to build their own SINIF according to their research purpose. More details about the workflow of the SINIF are explained in Sect. 4.1.

4 Methodology

The methodology for this study includes the review of the relevant literature and data collection from the NRC database. All NRC data during 1990 to 2017 were downloaded in the form of Excel files directly from the website of NRC (United States Coast Guard 2017). Each file contains several sheets with the official information (such as report time, call type, responsible company, and so on), incident information (such as description, time, location, cause, if it was a planned event, and so on), hazmat information (whether there was hazmat release, type of hazmat, amount of release, equipment type, and so on), weather conditions, and other information.

In this section, we explain the data processing method, the workflow of the SINIF, and the indices that are used to measure the performance of machine learning algorithms.

4.1 Data Processing and Analysis

The NRC data were analyzed according to the following steps, the summary of which can be considered as the workflow of SINIF:

- (1) The fields of “*SEQNOS*,” “*Description of incident*,” and “*Incident cause*” of all reports with a total number of 826,078 records between 1990 and 2017 in the NRC database were gathered into the total sample set named U.
- (2) It should be noted that a preliminary inspection of the sample set U led to an important observation related to this next step. In more detail, it seems that when the NRC call center received reports related to a previously recorded incident, a copy of the original report’s “*SEQNOS*” ID was also registered in the “*Description of incident*” field of the new report. Based on this finding, we eliminated 3641 identified duplicate reports from the sample set U by examining the “*Description of incident*” and “*SEQNOS*” fields respectively for multiple listings of the same “*SEQNOS*” ID, merging any repeating or fragmented descriptions of identified duplicate reports into one per incident.
- (3) Then the field “*Description of incident*” in U was analyzed through using the keyword extraction method with the aim of identifying and labeling the accident causes based on the same natural phenomenon classification used by Sengul et al. (2012) (that is, *Wind*, *Weather*, *Unknown*, *Tornado*, *Storm*, *Rain*, *Lightning*, *Hurricane*, *Flood*, *Earthquake*, *Cold*). Moreover, we decided to include an additional category named “NotNatech” so as to differentiate reports that are not Natech reports. After labeling all reports in set U using the selected keywords, we examined the content of the reports categorized as related to natural phenomena to assess if it was a planned event and whether there was hazmat release, in order to ensure those reports made reference to Natech reports. This process resulted in 32,348 records being identified as Natech reports and having the corresponding keyword label assigned to them, over 790,000 reports being labeled as “NotNatech.”
- (4) Based on the results of step 2, we manually checked the “*Description of incident*” of around 15,000 reports tagged with keywords related to natural phenomena and around 8000 reports labeled as “NotNatech” to ensure they had been labeled

correctly. Then these two groups of reports were merged as a new sample set named A with 24,060 records.

- (5) The LSTM and CNN algorithms were employed as the kernel networks to analyze set A, separately.
- (6) The sample set A was separated into a training set T_1 (80% of A) and a testing set T_2 (20% of A), so that the networks built in step 5 will be trained by T_1 and tested by T_2 accordingly. Afterwards, the text data in the T_1 dataset were transformed into sequences of word indices to match the requirements for the network training. A crucial point for the comparison of the two algorithms was to ensure that they were trained under the same conditions; therefore, we set the same values for the critical parameters for the two networks. For example, the learning rate was set at 0.01 and the epoch was set at 30.
- (7) The accuracy of the trained networks was then assessed and verified, and the network with the highest accuracy was selected to analyze the original sample set U.
- (8) Finally, a comparison of the results of steps 3 and 7 was carried out for each record to determine if the labels matched. If the results differed, the field “*Description of incident*” of each specific record was manually amended according to the researchers’ interpretation.

4.2 Performance Measurement Indices

For the purpose of measuring the performance of selected machine learning algorithms, we present several indicative measurements to verify the networks (Özgür et al. 2005; Sokolova and Lapalme 2009; Sboev et al. 2016). First, we separated T_2 into four subgroups for each individual class X_i shown below, and then we calculated the performance indices according to the size of each group for each individual class X_i .

True positive (TP): the item was labeled as class X_i in T_2 , and was also labeled as X_i by the network. False positive (FP): the item was not labeled as class X_i in T_2 , but was labeled as X_i by the network. True negative (TN): the item was not labeled as class X_i in T_2 , and was not also labeled as X_i by the network. False negative (FN): the item was labeled as class X_i in T_2 , and was not labeled as X_i by the network.

According to the above identified subgroups, we introduced three main performance indices to evaluate the accuracy of selected deep learning algorithm. The Precision (P_i) calculated by Eq. 1 was used to evaluate the class agreement of data labels with the labels given by the network for each individual class X_i .

$$P_i = \frac{TP_i}{TP_i + FP_i} \quad (1)$$

The Recall (R_i) in Eq. 2 indicates the effectiveness of selected algorithm to identify given labels for each individual class X_i .

$$R_i = \frac{TP_i}{TP_i + FN_i} \quad (2)$$

A harmonic average value ($F_{\beta=1}^i$) of Precision and Recall was used to assess the efficiency of selected algorithm for class X_i .

$$F_{\beta=1}^i = \frac{(1 + \beta^2)P_iR_i}{\beta^2 \cdot P_i + R_i}, \quad \beta = 1 \quad (3)$$

The Macro-average value ($\bar{F}$), which is the average value of $F_{\beta=1}^i$, was calculated by Eq. 4 to indicate the overall effectiveness of selected algorithm; the Accuracy (Acc) based on Eq. 5 was used to evaluate the overall accuracy of the algorithm.

$$\bar{F} = \frac{\sum_{i=1}^n F_i}{n} \quad (4)$$

$$\text{Acc} = \frac{\sum_{i=1}^n (TP_i + TN_i)}{\sum_{i=1}^n (TP_i + FP_i + FN_i + TN_i)} \quad (5)$$

5 Results and Discussion

In this section, we first focus on the total number of Natechs and the associated natural hazard causes for each Natech report in the NRC database. The analysis was conducted based on data generated through the keyword extraction method. Then, we proceeded to assess which machine learning algorithm of the two—the LSTM and CNN—was more suitable to implement the SINIF in order to retrieve Natech information from the NRC database. Finally, we offer a discussion on the quality of the SINIF classifying results for retrieving Natech data from the NRC database.

5.1 Keyword Extraction Results

According to the results of the keyword extraction method, the total number of Natechs reported to the NRC from 1990 to 2017 is 32,348, which represents 3.93% of all incidents reported to the NRC during that period and corresponds to 2.22% to 7.39% of the total incidents in each year. As shown in Fig. 2a, hurricanes produced the largest number of Natechs (22.73% in total), while storm triggered 20.60%, and an additional 38.71% were attributable to rain, wind, flood, and other undetermined weather hazards. Figure 2b explains the frequency of Natechs caused by

various natural phenomena in accordance with the results of the keyword extraction method. The total number of Natechs increased notably after 2005, due to the quantity change of hurricane-related Natechs. Meanwhile, another contributing factor is that hurricanes started to be registered as a separate triggering hazard from 2003 in the NRC database. Furthermore, it becomes apparent from Fig. 2a that using the keyword extraction method to identify the triggering natural hazard of Natechs has some issues, because there were 130 Natechs per year on average labeled as “*Unknown*.”

Although the results showed rates similar to those from the research of Sengul et al. (2012) considering the proportion of Natechs in all NRC incident reports for the study period, the triggering natural hazards demonstrate significant differences. The differences between these two results can probably be attributed to the following two aspects. First, the reports analyzed in this study cover a period from 1990 to 2017 instead of the period from 1990 to 2008, which was used in the original study of Sengul et al. (2012). Second, as mentioned in the introduction, due to the language expression, there will be a large bias if we only use the keyword retrieval method to analyze the NRC database without checking the description of an incident or checking other incident reports database. As a result, we can confirm that it is quite difficult to achieve the research target of this study through analyzing data by solely using the keyword extraction method.

5.2 Efficiency Analysis of Machine Learning Algorithms

In this study, we introduced machine learning to retrieve Natechs and identify their triggering natural hazard. We analyzed the training set and testing set by implementing two basic machine learning algorithms, the LSTM and the CNN, with the purpose of checking which algorithm is more suitable to achieve the research target. We then plotted the heat maps of the confusion matrix for each algorithm as shown in Fig. 3.

Each cell in the confusion matrix represents the number of records the network misclassified into the classes of their respective columns, when they should have been placed into the classes indicated by their rows. Thus, the confusion matrix describes the distribution of the predicted classes in respect to the classes that have been identified by the keyword extraction method.

Through the analysis of the confusion matrices, it becomes apparent that it is not easy to extract Natechs and identify their natural hazard causes by using only the machine learning method. Figure 3 shows that both the LSTM and the CNN are capable of identifying the probable triggering natural hazards for most of the Natech records.

Fig. 2 Number of Natech reports with various natural phenomena (a). Number of Natech reports associated with various natural phenomena according to the keyword extraction results (b). *Note:* the black line is the average number of Natech reports

However, due to the ambiguous descriptions of different natural hazards, there were certain biases in the Natechs triggered by hurricanes, storms, rain, or other weather-related causes. These findings are probably stemming from the fact that, for an untrained observer, the above natural hazards are induced by seemingly similar meteorological phenomena. Therefore, the vocabulary and expressions used to describe such accidents are often confusingly similar among Natechs. In addition, due to the same reason, some of the accidents, which were not triggered by natural hazards, were easily mistaken as Natechs by the networks. These ambiguities pose another set of challenges to this research endeavor, making it very difficult to

accomplish through the use of machine learning or keyword extraction alone. In this context and following the discussion in the introduction, the development and implementation of the SINIF can be considered a potential method to retrieve Natechs and identify the natural hazard cause successfully.

As explained in Sect. 3, we set out to find a suitable machine learning algorithm as the kernel network upon which the SINIF will be implemented. We calculated the indices explained in Sect. 4.2 in order to assess which machine learning algorithm is more suitable to accomplish the main task of this study. Table 2 presents the results of these performance indices for each algorithm respectively.

Fig. 3 Confusion matrix of classification results of the Long Short-Term Memory (LSTM) (a). Confusion matrix of classification results of the convolutional neural network (CNN) (b)

Table 2 Performance indices for each potential natural hazard cause determined by the machine learning algorithms

	Long Short-Term Memory (LSTM)			Convolutional Neural Network (CNN)		
	P_i	R_i	$F_{\beta=1}^i$	P_i	R_i	$F_{\beta=1}^i$
Cold	0.57	1.00	0.73	0.67	0.40	0.50
Earthquake	0.55	1.00	0.71	0.82	0.45	0.58
Flood	0.75	1.00	0.86	0.72	0.56	0.63
Hurricane	0.94	0.96	0.95	0.93	0.91	0.92
Lightning	0.96	0.93	0.94	0.91	0.90	0.91
NotNatech	0.95	0.93	0.94	0.86	0.95	0.90
Rain	0.97	0.93	0.95	0.77	0.75	0.76
Storm	0.95	0.93	0.94	0.80	0.75	0.77
Tornado	0.91	0.98	0.94	0.88	0.55	0.68
Weather	0.97	0.96	0.97	0.82	0.76	0.79
Wind	0.94	0.98	0.96	0.83	0.80	0.81

As the table shows, both the LSTM and the CNN scored differently depending on the index in each of the categories. In terms of precision indices for different categories, the LSTM could retrieve “NotNatech” and additionally identify the natural hazard cause more accurately than the CNN on Natechs triggered by flood, hurricane, lightning, rain, storm, tornado, weather, or wind. On

the other hand, from the viewpoint of recall indices, the CNN could retrieve more “NotNatech” than the LSTM. As a comprehensive index of precision and recall, the harmonic average ($F_{\beta=1}^i$) index shows that the LSTM is superior to the CNN, a difference ranging from 0.03 up to 0.26 in favor of the LSTM was observed across all categories (Table 2). Furthermore, we calculated the macro-

Table 3 Examples of National Response Center (NRC) reports identified by the Semi-Intelligent Natech Identification Framework (SINIF)

Description of incident ^d	Type of incident ^a	Incident cause in the NRC database	SINIF labels	CHRIS code ^{b,c}	Released
Below ground and water 4IN dot ^e regulated pipeline/while flushing line in an attempt to repair damages caused by Hurricane George	Platform	Equipment failure	hurricane	OIL	Yes
A corrosive liquid material spilled onto the deck of a container vessel from a 55 gallon drum. Due to heavy rain the some material has washed into the Atlantic Ocean. The vessel is bound for hampton roads, virginia on morning of 10 SEP 2002	Vessel	Equipment failure	Rain	NCC	Yes
Aton ^f battery release report received via coast guard message traffic	Vessel	Hurricane	NotNatech	MAR	Yes
Caller stated that at 13:00 (E.D.T.) the Diablo Canyon power plant notified the U.S. nuclear regulatory commission of an unusual event at their facility. The event was brought on by an earthquake measuring between 6.0–7.0 on the richter scale and generating a vertical displacement of 0.012 g	Fixed	Earthquake	NotNatech	RAM	No
Caller stated a crushed steel storage tank was blown in from Hurricane Ike and has the partial number of 326 on the tank the forth number can not be seen. The tank has not been inspected or touched by anyone due to the corrosive materials on the storage tank. Caller estimated the size of the tank capacity at 4000 gallons	Storage tank	Hurricane	Hurricane	UNK	Yes
Caller reported that due to the tropical storm the water rose above the salt water barge which is fixed and residual oil from the sump pump caused a sheen on the water	Fixed	Tornado	Hurricane	OWA	Yes

^aIn the type of incident field, “Fixed” means that this incident happened in a fixed installation

^bDue to no cause-mentioned description or no chemical material released during the incident, labeled as “NotNatech” report

^cChemical Hazard Response Information System (CHRIS) code is used in the NRC database to determine the type of released materials

^dAll the descriptions of incident, type of incident, and incident cause in the NRC database listed here were copied from the NRC database directly. We did not make any modification on spelling or grammar

^eDepartment of Transportation (DOT), United States

^fAton is a registered company to produce batteries. <http://www.aton-energy.com/>

average value and accuracy for both algorithms under examination. The results show that the LSTM managed to retrieve 94.28% Natechs from the testing set, while the CNN only 85.64%. A larger macro-average value (0.8979) meant that the LSTM was more accurate than its competitor algorithm, which achieved a value of just 0.7497. The latter two indices emphatically show that the LSTM performed better in general and thus, is more suitable than the CNN in achieving the research target of this study. Moreover, the accuracy values of the LSTM and the CNN show that machine learning theory and the proposed SINIF in particular are more than capable in retrieving Natechs and identifying the associated natural hazards.

There are two key reasons explaining the effectiveness of the LSTM over the CNN in extracting Natechs from the NRC database. First, the NRC data were organized as time series data in this study. This is supported by LeCun and Bengio (1997), who noted that CNN is not as competent as the recurrent neural networks (RNN) in analyzing time series data, while the LSTM is one advanced type of RNN. Second, the aim of this study has been framed as a kind of text multi-classification problem, and as LeCun et al.

(2015) suggested, the LSTM algorithm is better suited to solve text-learning problems.

5.3 Semi-Intelligent Natech Identification Framework (SINIF) Results

The arguments supporting why LSTM is more suitable than CNN in designing the identification framework for the extraction of Natechs from the NRC database were presented above. After developing the SINIF and preparing the input dataset by removing any duplicates, the remaining NRC records between 1990 and 2017 (822,437 reports in total) were examined by the trained LSTM network. Then, they were classified either according to the triggering natural hazard, receiving a corresponding label, or as “NotNatech” if the hazmat release report was not caused by a natural hazard. The classification results of the LSTM network were then compared with the results of the keyword extraction method. There were 791,085 records successfully tagged either as “NotNatech” or with the corresponding label using the LSTM network; the keyword extraction method yielded the same record number. Of

Fig. 4 Number of Natech reports with various natural phenomena (a). Number of Natech reports associated with various natural phenomena according to SINIF results (b). Note: the black line is the average number of Natech reports

these, 25,467 records were classified as Natech reports. They were grouped as “certain Natech reports” set R_C . In contrast, the other 31,352 records did not obtain matching labels from the two methods, and thus formed the group “uncertain Natech reports” set R_U . After manually rechecking the field “Description of incident” and “Incident cause” for each record of R_U , an extra 7444 accidents were identified and added to the final set of Natech reports R_N .

Although we attempted to reduce duplicate data and succeeded to a certain extent on step 2, following the

method described in Sect. 4.1, there may still be other types of duplicate reports within R_N .

The NRC call center receives and catalogues hazmat-release reports made by any citizens in the United States. Due to the lack of information and multiple report sources, the NRC call center could not ascertain whether these reports were related to an already registered report or not. For the propose of reducing this kind of duplicate reports, we had to assume that all the information for the chemical release incidents recorded in the NRC database is correct, while the remaining doubt pertains to whether they were actually related to a single incident or not. Based on this,

1990 to 2017. These Natech reports comprise 3.98% of all incident reports in the NRC database, and between 2.26% and 7.27% of the total reports each year. Figure 4a shows that hurricanes were mentioned in the largest number of Natech reports (24.42%), while 19.27% and 18.29% of the Natech reports included references of “rain” and “storm” respectively. An additional 36.26% was attributable to wind, flood, cold, tornado, and other weather-related events. Figure 4b shows that the number of Natech reports has an upward trend from 1990 to 2017. Apart from 2016, the number of Natech reports has exceeded the average value after 2005, 2008, and 2012 respectively, which coincides with a higher report count of hurricane-induced Natechs. The number of hurricane-triggered Natech reports increased sharply and fluctuated drastically after 2003, whereas the frequencies of rain, storm, flood, and wind caused Natech reports remained relatively stable for the same time frame.

During the process of comparing the results of the SINIF with those from the keyword extraction method alone, Natech reports involving snow, heat, tide, rough sea, and mudslides were also retrieved. Such records comprised 3.01% of the total number of the Natech reports in the NRC database during the study period. This finding shows that the accuracy of the keyword extraction method is directly affected by the selection of keywords. If the selected keywords do not explicitly include terms that describe the potential natural hazard types, the process is highly probable to miss reports of Natechs triggered by natural hazards not defined in advance. In other words, the system has no information about what to look for initially, and so fails to identify the triggering natural hazards. However, due to the comparison function embedded in SINIF’s design, the framework is able to identify potential natural hazards, which might otherwise have been omitted by the keyword extraction method alone. By complementarily employing a deep learning method through the neural networks embedded in its structure, the SINIF actually attempts to identify the underlying patterns in the descriptions of the records in order to determine the triggering natural hazard, instead of seeking predefined terms in the text. Therefore, the SINIF seems to hold an advantage over the sole application of the keyword extraction method in this regard, as it offers an added identification and verification step.

Another point to note is the distribution of Natech reports with regard to various climatic changes and natural phenomena. Generally, the distribution patterns generated by the SINIF seem more reasonable than the ones from the keyword extraction method in the sense that they align with certain observed environmental tendencies. Several researchers have noted a dramatic rise in terms of intensity, size, duration, and frequency of hurricanes around the

Atlantic Ocean since the early 1980s, which could be attributed to climate change (Emanuel 2007; Knutson et al. 2010; Torn and Snyder 2012; Landsea and Franklin 2013). It is not unreasonable to assume, consequently, that hurricanes and hurricane-related natural hazards, such as storms, rains, and floods, are becoming increasingly likely to trigger Natechs, and that this tendency will continue in the future. Figure 4b suggests that the number of Natech reports related to hurricanes grew remarkably since 2003, with notable spikes in 2005 when Rita, Wilma, and Katrina struck, 2008 with Ike, 2012 with Sandy, and 2017 when Harvey, Maria, and Irma occurred. Indeed, hurricane-related Natech reports increased 16-fold since the year 2003, thereby supporting this alarming trend. Apart from the above, the terms most commonly used in the incident descriptions were plotted in the form of a word cloud (Fig. 5), in order to evaluate their frequency throughout the NRC records. As expected, hurricane, storm, and rain were almost always mentioned in the descriptions, they are indeed the most reported triggering cause as evidenced by Fig. 5. This means that the majority of Natech reports throughout the period under investigation were probably induced by such weather-related causes. The fact that the SINIF demonstrated satisfactory competency in handling the practical issues and thus bringing into light these findings, serves as another example to support that it is a more suitable and accurate tool compared to the sole application of the keyword extraction method in retrieving Natech reports and identifying their associated natural hazards.

6 Conclusion

This study developed a Semi-Intelligent Natech Identification Framework (SINIF) to retrieve Natech related reports and identify the associated triggering natural hazards from the NRC database. The incident reports recorded in the NRC database between 1990 and 2017 were analyzed through the keyword extraction method and the SINIF, separately. The comparison of the results for the keyword extraction method alone and the SINIF indicates that the latter is a more suitable analytical technique to achieve accurately and efficiently the research purpose, namely the extraction of Natechs from a large database (the NRC database in this study). First, compared to using the keyword extraction method alone, keyword selection in the SINIF plays by default a smaller role in determining the extraction results, since it is combined with an additional structure of a deep learning network. Second, the SINIF’s results are more reasonable compared to those from the keyword extraction method alone. Third, the SINIF is capable of analyzing large databases with comparatively

higher accuracy than the alternative method of keyword extraction.

According to the results of the SINIF, 3.98% of hazmat release accidents were identified as Natech-related reports between 1990 and 2017. Furthermore, according to the retrieved NRC reports, the majority of Natech reports (98.24%) were mentioned as related to meteorological phenomena, while hurricanes (24.42%), heavy rains (19.27%), and storms (18.29%) were detected as the main causes of Natech reports. As evidenced by the findings of this study, the observed rising overall number of Natech reports since 2004 can be primarily attributed to the increased occurrence of hurricane-triggered Natechs. The frequency and severity of extreme weather-related events, such as hurricanes, might rise due to climate change. In turn, such an increase in these parameters would suggest that Natech reports could become more likely to happen in the future. However, due to the uncertainty of Natech events, we should be careful to conclude that the increasing trend on the number of Natech reports or events could be attributed to climate change. But the findings of this study points to an interesting research direction on whether climate change has an impact on the incidence of Natech event, and how it works.

As with any data extraction method, the human factor poses significant limitations to the SINIF. Arguably, its framework is very dependent on the keyword extraction method by design. The fact that this technique relies so heavily on the initial selection of searching terms from the researchers, as demonstrated in this study as well, introduces the human error in this machine learning framework. Even the network analysis in steps 4 through 7 and the comparison in step 8 are incapable of completely eliminating such effects. According to the workflow of the SINIF, in order to retrieve the target records from large databases, researchers must input “reasonable” search keywords—terms that they estimate adequately describe the target incidents—into the SINIF. However, no matter how carefully the researchers select keywords, the risk that the selected keywords might not cover all the potential Natech reports still exists unless they manually check millions of chemical release records within the large database. Furthermore, if the researchers’ selected keywords are not included in the actual incident descriptions within the dataset, such records may not be selected to be added to the training data in the following step, because the SINIF has no way of assessing the keywords themselves. This may even result in a number of Natech reports within the database not being correctly identified as such. Another shortcoming of the SINIF is related to the selection of machine learning algorithm, which is implemented as the kernel network. As demonstrated by our analysis, various algorithms can produce starkly different results in

classifying the descriptions of accidents. Therefore, we recommend researchers who want to apply the SINIF to assess the respective advantages and disadvantages of each kernel algorithm beforehand. In our case, the LSTM proved more suitable than the CNN to build the SINIF for retrieving Natech reports from the NRC database. The last weakness of the SINIF is the high requirements of technical expertise in computer science. The users should have a basic understanding of the machine learning algorithm they select in order to modify accordingly the parameters of the kernel network, and assure the accuracy of their results. In other words, the SINIF is not quite user-friendly yet.

Despite the aforementioned disadvantages of the SINIF, there are also significant benefits for researchers dealing with technological accidents. This endeavor set out to develop a machine learning algorithm for the extraction of Natech related reports from the NRC database and the identification of their triggering natural hazard causes. However, an important asset of this SINIF is its versatility. Naturally, this framework can be applied to analyze a wide array of databases, such as the FACTS or eMARS. Minor modifications may be necessary to take into account the different field names used in each database, though. But apart from employing the SINIF in order to accomplish the same task with different initial datasets, the algorithm can easily be used by researchers to focus on specific categories of Natechs, or even to examine other aspects of technological accidents depending on the information available in the dataset. In addition, because of its high compatibility, users can implement the SINIF based on any of the existing machine learning algorithms to improve the accuracy of the extraction results.

Acknowledgements This research was supported in part by the Japan Society for the Promotion of Science (Kaken Grant 17K01336, April 2017–March 2020). We would also like to acknowledge that Xiaolong Luo holds a Ph.D. fellowship from the China Scholarship Council (CSC No. 201606620007) and the Ministry of Education, Culture, Sports, Science, and Technology of Japan (Monbukagakusho: MEXT Scholarship No. 171572, 2017–2021), and Dimitrios Tzioutzios holds a Ph.D. fellowship from the Ministry of Education, Culture, Sports, Science, and Technology of Japan (Monbukagakusho: MEXT scholarship, 2019–2022).

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright

holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Antonioni, G., S. Bonvicini, G. Spadoni, and V. Cozzani. 2009. Development of a framework for the risk assessment of Na-Tech accidental events. *Reliability Engineering and System Safety* 94(9): 1442–1450.
- Antonioni, G., G. Landucci, A. Necci, D. Gheorghiu, and V. Cozzani. 2015. Quantitative assessment of risk due to NaTech scenarios caused by floods. *Reliability Engineering and System Safety* 142: 334–345.
- Antonioni, G., G. Spadoni, and V. Cozzani. 2007. A methodology for the quantitative risk assessment of major accidents triggered by seismic events. *Journal of Hazardous Materials* 147(1–2): 48–59.
- Bashar, S.S., and A.A. Mahmud. 2019. A machine learning approach for heart rate estimation from PPG signal using random forest regression algorithm. In *Proceedings of 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, 24–25 July 2019, Swat, Pakistan, 1–5.
- Bishop, C.M. 2006. *Pattern recognition and machine learning*. Berlin: Springer.
- Blanzieri, E., and A. Bryl. 2008. A survey of learning-based techniques of email spam filtering. *Artificial Intelligence Review* 29: 63–92.
- Bureau for Analysis of Industrial Risk and Pollution. 2019. Analysis, Research and Information on Accidents (ARIA). French Ministry of Ecology and Sustainable Development, Bureau for Analysis of Industrial Risk and Pollution, France. <https://www.aria.developpement-durable.gouv.fr>. Accessed 10 Jan 2019 (in French).
- Campadelli, P., R. Lanzarotti, and G. Lipori. 2004. Face detection in color images of generic scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24: 97–103.
- Campedel, M., V. Cozzani, A. Garcia-Agreda, and E. Salzano. 2008. Extending the quantitative assessment of industrial risks to earthquake effects. *Risk Analysis* 28: 1231–1246.
- Chowdhury, G.G. 2003. Natural language processing. *Annual Review of Information Science and Technology* 37: 51–89.
- Cireşan, D., U. Meier, and J. Schmidhuber. 2012. Multi-column deep neural networks for image classification. *arXiv preprint arXiv:1202.2745*.
- Clow, J.C. 1980. The Coast Guard's National Response Center. *Environment International* 3(2): 151–153.
- Cozzani, V., M. Campedel, E. Renni, and E. Krausmann. 2010. Industrial accidents triggered by flood events: Analysis of past accidents. *Journal of Hazardous Materials* 175: 501–509.
- Cruz, A.M., and E. Krausmann. 2009. Hazardous-materials releases from offshore oil and gas facilities and emergency response following Hurricanes Katrina and Rita. *Journal of Loss Prevention in the Process Industries* 22(1): 59–65.
- Cruz, A.M., and N. Okada. 2008. Consideration of natural hazards in the design and risk management of industrial facilities. *Natural Hazards* 44(2): 213–227.
- Cruz, A.M., L.J. Steinberg, and A.L. Vetere-Arellano. 2006. Emerging issues for natech disaster risk management in Europe. *Journal of Risk Research* 9(5): 483–501.
- Diale, M., T. Celik, and C. Van Der Walt. 2019. Unsupervised feature learning for spam email filtering. *Computers and Electrical Engineering* 74: 89–104.
- Emanuel, K. 2007. Environmental factors affecting tropical cyclone power dissipation. *Journal of Climate* 20(22): 5497–5509.
- European Commission. 2019a. Natech accident database. Joint Research Centre, Institute for the Protection and Security of the Citizen, Italy. <http://enatech.jrc.ec.europa.eu>. Accessed 11 Jan 2019.
- European Commission. 2019b. eMars (Major Accident Reporting System) database. European Commission, Joint Research Centre, Institute for the Protection and Security of the Citizen, Italy. <http://emars.jrc.ec.europa.eu>. Accessed 11 Jan 2019.
- Fernández, S., A. Graves, and J. Schmidhuber. 2007. An application of recurrent neural networks to discriminative keyword spotting. In *Proceedings of International Conference on Artificial Neural Networks*, 9–13 September 2007, Porto, Portugal, 220–229.
- Fukushima, K. 1980. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics* 36(4): 193–202.
- Girgin, S., and E. Krausmann. 2014. Analysis of pipeline accidents induced by natural hazards: Final report. JRC88410. Joint Research Centre, European Commission, Italy.
- Girgin, S., and E. Krausmann. 2016. Historical analysis of U.S. onshore hazardous liquid pipeline accidents triggered by natural hazards. *Journal of Loss Prevention in the Process Industries* 40: 578–590.
- Graves, A., and J. Schmidhuber. 2009. Offline handwriting recognition with multidimensional recurrent neural networks. In *Proceedings of the 21st International Conference on Neural Information Processing Systems*, ed. D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, 545–552. Red Hook, NY: Curran Associates Inc.
- Hochreiter, S., and J. Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9(8): 1735–1780.
- Huang, G.-B., H. Zhou, X. Ding, and R. Zhang. 2011. Extreme learning machine for regression and multiclass classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 42(2): 513–529.
- Ji, S., W. Xu, M. Yang, and K. Yu. 2013. 3D convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35(1): 221–231.
- Khakzad, N., and P. Van Gelder. 2017. Fragility assessment of chemical storage tanks subject to floods. *Process Safety and Environmental Protection* 111: 75–84.
- Khakzad, N., and P. Van Gelder. 2018. Vulnerability of industrial plants to flood-induced natechs: A Bayesian network approach. *Reliability Engineering and System Safety* 169: 403–411.
- Khakzad, N., M. Dadashzadeh, and G. Reniers. 2018. Quantitative assessment of wildfire risk in oil facilities. *Journal of Environmental Management* 223: 433–443.
- Knutson, T.R., R.E. Tuleya, S.T. Garner, M.A. Bender, G.A. Vecchi, I.M. Held, and J.J. Sirutis. 2010. Modeled impact of anthropogenic warming on the frequency of intense Atlantic hurricanes. *Science* 327(5964): 454–458.
- Krausmann, E., and A.M. Cruz. 2013. Impact of the 11 March 2011, Great East Japan earthquake and tsunami on the chemical industry. *Natural Hazards* 67(2): 811–828.
- Krausmann, E., and F. Mushtaq. 2008. A qualitative Natech damage scale for the impact of floods on selected industrial facilities. *Natural Hazards* 46(2): 179–197.
- Krausmann, E., and E. Salzano. 2017. Lessons learned from Natech events. In *Natech risk assessment and management: Reducing the risk of natural-hazard impact on hazardous installations*, ed. E. Krausmann, A.M. Cruz, and E. Salzano, 33–52. Amsterdam: Elsevier.
- Krausmann, E., E. Renni, M. Campedel, and V. Cozzani. 2011. Industrial accidents triggered by earthquakes, floods and lightning: Lessons learned from a database analysis. *Natural Hazards* 59(1): 285–300.

- Kumasaki, M., T. Hara, N. Nakajima, Y. Wada, and R. Makino. 2017. The classification of physical effects from natural hazards for Natech risk assessment based on a Japanese database. *Journal of Loss Prevention in the Process Industries* 50(B): 308–316.
- Landsea, C.W., and J.L. Franklin. 2013. Atlantic hurricane database uncertainty and presentation of a new database format. *Monthly Weather Review* 141(10): 3576–3592.
- Le, Q.V., and T. Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning*, 21–26 June 2014, Beijing, China. JMLR:W&CP 32: 1–5.
- LeCun, Y., and Y. Bengio. 1997. Convolutional networks for images, speech, and time series. In *The handbook of brain theory and neural networks*, ed. M.A. Arbib, 276–279. Cambridge, MA: The MIT Press.
- LeCun, Y., Y. Bengio, and G. Hinton. 2015. Deep learning. *Nature* 521(7553): 436–444.
- Lu, D., and Q. Weng. 2007. A survey of image classification methods and techniques for improving classification performance. *International Journal of Remote Sensing* 28(5): 823–870.
- Manevitz, L.M., and M. Yousef. 2001. One-class SVMs for document classification. *Journal of Machine Learning Research* 2: 139–154.
- Murphy, K.P. 2012. *Machine learning: A probabilistic perspective*. Cambridge, MA: The MIT press.
- Özgür, A., L. Özgür, and T. Güngör. 2005. Text categorization with class-based and corpus-based keyword selection. In *Proceedings of 20th International Symposium on Computer and Information Sciences*, 26–28 October 2005, Istanbul, Turkey, 606–615.
- Ranjan, R., V.M. Patel, and R. Chellappa. 2019. HyperFace: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41: 121–135.
- Rubin, T.N., A. Chambers, P. Smyth, and M. Steyvers. 2012. Statistical topic models for multi-label document classification. *Machine Learning* 88(1–2): 157–208.
- Sboev, A., T. Litvinova, D. Gudovskikh, R. Rybka, and I. Moloshnikov. 2016. Machine learning models of text categorization by author gender using topic-independent features. *Procedia Computer Science* 101: 135–142.
- Sengul, H., N. Santella, L.J. Steinberg, and A.M. Cruz. 2012. Analysis of hazardous material releases due to natural hazards in the United States. *Disasters* 36(4): 723–743.
- Shah, A.A., J. Ye, M. Abid, J. Khan, and S.M. Amir. 2018. Flood hazards: Household vulnerability and resilience in disaster-prone districts of Khyber Pakhtunkhwa Province, Pakistan. *Natural Hazards* 93(1): 147–165.
- Shin, J., Y. Kim, S. Yoon, and K. Jung. 2018a. Contextual-CNN: A novel architecture capturing unified meaning for sentence classification. In *Proceedings of 2018 IEEE International Conference on Big Data and Smart Computing*, 15–18 January 2018, Shanghai, China, 491–494.
- Shih, C.H., B.C. Yan, S.H. Liu, and B. Chen. 2018b. Investigating Siamese LSTM networks for text categorization. In *Proceedings of the 9th Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 12–15 November 2018, Honolulu, Hawaii, USA, 641–646.
- Showalter, P.S., and M.F. Myers. 1992. Natural disasters as the cause of technological emergencies: A review of the decade, 1980–1989. Boulder, CO: University of Colorado.
- Sokolova, M., and G. Lapalme. 2009. A systematic analysis of performance measures for classification tasks. *Information Processing and Management* 45(4): 427–437.
- Sotthisopha, N., and P. Vateekul. 2018. Improving short text classification using fast semantic expansion on multichannel convolutional neural network. In *Proceedings of the 19th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*, 27–29 June 2018, Busan, South Korea, 182–187.
- Sudharshan, P.J., C. Petitjean, F. Spanhol, L.E. Oliveira, L. Heutte, and P. Honeine. 2019. Multiple instance learning for histopathological breast cancer image classification. *Expert Systems with Applications* 117: 103–111.
- TNO Industrial and External Safety. 2019. Failure and Accidents Technical information System (FACTS). the Unified Industrial & Harbour Fire Department in Rotterdam-Rozenburg, the Netherlands. <http://www.factsonline.nl/>. Accessed 14 Jan 2019.
- Torn, R.D., and C. Snyder. 2012. Uncertainty of tropical cyclone best-track information. *Weather and Forecasting* 27(3): 715–729.
- United States Coast Guard. 2017. United States National Response Center (NRC) database. Washington, DC: United States Coast Guard. <http://www.nrc.uscg.mil/>. Accessed 17 Oct 2017.