Article

A C-BiLSTM Approach to Classify Construction Accident Reports

Jinyue Zhang [1],[†] , Lijun Zi [2],[†], Yuexian Hou [3],[*],[†], Da Deng [3], Wenting Jiang [2] and Mingen Wang [3]

1   College of Management and Economics, Tianjin University-Trimble Joint Laboratory for BIM,
     Tianjin University, 92 Weijin Road, Tianjin 300072, China; jinyuezhang@tju.edu.cn
2   Guangzhou Metro Design and Research Institute Co. Ltd., 204 Huanshi West Road,
     Guangzhou 510010, China; zilijun@dtsjy.com (L.Z.); jiangwenting@dtsjy.com (W.J.)
3   College of Intelligence and Computing, Tianjin University, 92 Weijin Road, Tianjin 300072, China;
     dengda@tju.edu.cn (D.D.); wangmingen@tju.edu.cn (M.W.)
*   Correspondence: yxhou@tju.edu.cn; Tel.: +86-22-2740-2272
†   Jinyue Zhang, Lijun Zi, and Yuexian Hou have equal contribution to this paper.

check for updates

Abstract: The construction sector is widely recognized as having the most hazardous working environment among the various business sectors, and many research studies have focused on injury prevention strategies for use on construction sites. The risk-based theory emphasizes the analysis of accident causes extracted from accident reports to understand, predict, and prevent the occurrence of construction accidents. The first step in the analysis is to classify the incidents from a massive number of reports into different cause categories, a task which is usually performed on a manual basis by domain experts. The research described in this paper proposes a convolutional bidirectional long short-term memory (C-BiLSTM)-based method to automatically classify construction accident reports. The proposed approach was applied on a dataset of construction accident narratives obtained from the Occupational Safety and Health Administration website, and the results indicate that this model performs better than some of the classic machine learning models commonly used in classification tasks, including support vector machine (SVM), naïve Bayes (NB), and logistic regression (LR). The results of this study can help safety managers to develop risk management strategies.

1. Introduction

Workplace health and safety is a significant concern in all countries [1] because there are more than 2.78 million deaths caused by occupational accidents every year according to the International Labour Organization [2]. The construction industry is recognized as the most hazardous one among various industries [3]. In the United States, construction accounts for approximately one-sixth of fatal accidents while only employing 7% of the national workforce, and there are four recorded injuries per 100 full-time construction workers in the construction industry.

Construction accidents usually result in both health/safety issues and financial loss [4], and thus there has been abundant research motivated by the alarming injury and fatality rates. Research on construction safety is mainly conducted from two perspectives: it is either management-driven or technology-driven [5]. In general, it is assumed that enhanced construction safety management can effectively improve on-site safety performance and reduce the number of accidents. Research from the management perspective usually includes either safety management processes such as safety education and training or focuses on individual/organizational characteristics such as workers' attitudes towards safety. However, the effect of traditional strategies for preventing injuries was limited due to their

reactive and regulatory-based nature [6]. Esmaeili and Hallowell [7] indicate that the construction industry has reached saturation with respect to these injury prevention strategies. Along with the advancement of information and communication technology, various innovative technologies have been investigated to assist and improve on existing management-driven safety management practices. These technical approaches aimed to enhance rather than replace management efforts [8].

Besides the assistance of technologies, some new injury prevention strategies have been developed for the construction industry. The risk analysis method is one of them which are used in safety programs to improve safety performance. For example, Baradan and Usmen [9] compared the risk of different building trades, Hallowell and Gambatese [10] quantified the safety risk for various activities required to construct concrete formwork, and Shapira and Lyachin [11] studied the impact of tower cranes on job site safety. However, most of these risk-based studies are limited to specific application fields and hard to translate well to a general scope of the construction industry.

To address this limitation in the previous literature, Esmaeili and Hallowell [12] proposed an attribute-based risk identification and analysis method that helps designers to identify and model the safety risk independently of specific activities or trades. In this method, accidents are considered the outcome of interaction among physical conditions of the jobsite, environmental factors, administrative issues, and human error. Although this method shows promise, it requires the analysis of large numbers of construction injury reports to first classify the causes and then see patterns and trends that emerge from the data. Such manual content analysis is laborious and resource-intensive [13].

It is vitally important to analyze past accidents and understand the causes to prevent the occurrence of similar accidents and promote workplace safety [14] by removing or reducing the identified causes. Construction injury reports contain a wealth of empirical knowledge that could be used to better understand, predict, and prevent the occurrence of construction accidents. Some major construction companies and federal agencies, for example, the Occupational Safety and Health Administration (OSHA), possess those reports in the form of huge digital databases. Because different companies have different requirements for the forms of accident reports, these reports are often unstructured or semi-structured.

The first step towards the effective analysis of construction injury reports is the rigorous classification based on accident causes. In the construction industry, early text classification was achieved manually, which is not only demanding for professional knowledge but also requires a large amount of human and material resources. Furthermore, the consistency of the classification results is difficult to be ensured. Therefore, it is significantly important to investigate the method of automatic classification of texts written in natural language [15]. However, studies of text mining, natural language processing (NLP) and deep learning (DL) techniques for the analysis of construction accident narratives are very limited [16]. To fill this gap, using accident narratives data obtained from the official website of OSHA, this paper presents a novel and unified architecture that contains a bidirectional long short-term memory (BiLSTM) model and a convolutional layer for the classification of construction accident causes. The proposed architecture is called convolutional-BiLSTM (C-BiLSTM). This novel construction accident report classification model was compared with some advanced methods in previous research work using a set of OSHA data and indicated a superior result than other models.

The rest of this paper is organized as follows. Related works are presented in the next Section, including text mining and machine learning techniques, existing studies on accident narrative classification, and performance metrics. Then, the research approach is presented in detail in Section 3, along with the method for data pre-processing. Before the conclusion section, the study discusses the result of applying this introduced approach to the OSHA accident narratives and compares its performance with the state-of-the-art approaches in text classification.

2. Related Works

2.1. Text Mining and Machine Learning

Text mining refers to obtain valuable information and knowledge from text data, which is a method in data mining [17]. With the fast coming of big data, the use of massive data has been reforming every industry and business, becoming an important production factor [18]. In many cases, the text data is one of the most easily generated data forms, with typical features of unstructured data. Although the unstructured text is easily perceived and handled by humans, it is hard to be understood by machines. The most basic yet important application of text mining is to achieve automatic classification based on text content [19]. Text classification is one of the common tasks in NLP which concerns the method to program computers to process and analyze large amounts of natural language data. In-text classification, a mathematical model is trained by a set of input texts with associated classification tags to have certain generalization ability so that the model can perform a good prediction on the category of other texts in the same domain. It is essential to measure and calculate the similarity of texts.

Currently, as an effective method for text information management, text classification has been widely applied to multiple fields such as information classification [20], recommendation system [21], and sentiment analysis [22]. Traditional text classification methods generally adopt a machine learning method [23]. Machine learning is the scientific study of algorithms and statistical models to perform a specific task without using explicit instructions but relying on patterns and inference. The mathematical model is built based on sample data (as known as training data) to make predictions. Various types of models have been used and researched for machine learning systems, including Support Vector Machine (SVM) [24], Naive Bayes (NB) [25], K neighbors [26], etc.

DL is a class of machine learning algorithms that uses multiple layers to progressively extract higher-level features from the raw input [27]. For example, in image processing, lower layers may identify edges, while higher layers may identify the concepts relevant to a human such as faces. DL method has been proved to be effective for feature extraction [28]. A series of DL algorithms such as Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) have been extensively used by researchers in various fields for text classification [29,30].

The transformer architecture is proposed to deal with the difficulty of parallel training related to BiLSTMs. It completely replaces LSTMs by the so-called attention mechanism [31]. With attention, an entire sequence is treated as a whole, therefore it is much easier to train in parallel. There are many variants of attention mechanisms, such as co-attention mechanism [32] and self-attention [33]. As a transformer-based approach, BERT (Bidirectional Encoder Representations from Transformers) [34] has achieved amazing results in many language understanding tasks, including the tasks of text classification. However, those advanced models usually have a large size with many parameters, making a higher cost in training.

2.2. Existing Studies on Accident Narrative Classification

There are some existing studies in the field of accident classification using machine learning approaches. Bertke et al. [35] used an NB-based model to classify the reasons for insurance claims on work-related injuries. The overall accuracy of the model is approximately 90%, but the accuracy of the category of claim on minor injury is somehow decreased. Tanguy et al. [36] evaluated the aviation safety reports by an SVM-based model and obtained an accuracy rate of 60–96%. Wellman et al. [37] proposed a fuzzy Bayesian model to classify the injury reports obtained from the National Health Interview Survey (NHIS) and achieved an accuracy rate of 87%. Abdat et al. [38] extracted the scenes of high recurrence Occupational Accident with Movement Disturbance (OAMD) from the narrative texts using a Bayesian network. However, this method requires expert knowledge to pre-process data and it is time-consuming. Zhong et al. [29] classified building quality complaint reports using a CNN model, and the weighted average of F1 values is 0.73, which is superior to the results of traditional machine learning algorithms such as NB and SVM.

In the application field of construction accident classification, related studies are very limited. Tixier et al. [39] proposed a rule-based automatic content analysis system that automatically extracts attributes and safety outcomes from unstructured injury reports. This system achieved an accuracy rate of 95%, but it showed poor performance when dealing with unexpected situations. Moreover, this approach requires an external dictionary for professional terminologies. Goh [40] used the data obtained from OSHA website to compare the performance of several machine learning algorithms, including SVM, NB, decision trees (DT), linear regression (LR), random forest (RF) and k-nearest neighbor (KNN), in the classification of construction accident reports. The results showed that the SVM-based classifier generated a better F1 value than other classifiers. Zhang et al. [16] further proposed sequential quadratic programming (SQP)-based integrated algorithm based on Goh et al.'s work. This combined method achieved a weighted result of 0.68, which is better than the result of a single machine learning algorithm.

It is not difficult to find that although there are some studies on the classification of construction accidents by traditional machine learning algorithms, there is still a lack of research on the application of DL algorithms in this field. Therefore, this study is aimed to evaluate the performance of the DL algorithm in the automatic classification of construction accident narratives.

2.3. LSTM, BiLSTM, and C-BiLSTM

In recent years, LSTM has been applied more widely. To further improve the performance of LSTM in processing variable-length sequence information tasks, researchers have proposed many methods to improve LSTM. The combination of LSTM or its variants with other network structures is an important research direction at present. Lu et al. [41] proposed a new emotion classification model called P-LSTM. By introducing a phrase factor mechanism, the P-LSTM model can extract more accurate information from text. Wang et al. [42] used a BiLSTM model to perform sequence analysis of microblog conversations to capture the distance dependence of the emotional semantic field. Experiments show that the BiLSTM model with context information is superior to other algorithms. Wei et al. [43] proposed a migration learning framework, ConvL, based on CNN and LSTM, which was used to automatically identify whether online comments expressed confusion, determine the degree of urgency and classify the polarity of emotions. Le et al. [44] introduced multi-View recurrent neural networks (MV-RNN) for 3D network segmentation. This framework combines CNN and double-layer LSTM for 3D shape segmentation, which can output the edge image of each defined view. Harish et al. [45] used the model combining CNN and BiLSTM to automatically identify inappropriate query suggestions, and the performance of this model is better than that of multiple benchmark models using the same data set for training. The model used in this study is based on the above research and further improved according to the research needs. The specific content of the model will be highlighted in the next chapter.

2.4. Performance Metrics

In most existing studies, the F1 score proposed by Buckland and Gey [46] was widely used as a performance indicator to evaluate the classification model. This indicator considers both the Precision and the Recall of the test to compute the F1 score: Precision is the number of correct positive results divided by the number of all positive results returned by the classifier, and Recall is the number of correct positive results divided by the number of all relevant samples (all samples that should have been identified as positive). The F1 score is equivalent to the comprehensive evaluation indicator of Precision and Recall. Equation (1) lists the calculations for Precision, Recall, and F1 scores respectively.

Where TP refers to the number of positive samples that are classified correctly, FP refers to the number of negative samples that are classified to be positive, and FN refers to the number of positive samples that are classified to be negative.

$$recall = \frac{TP}{TP+FN}$$
$$precision = \frac{TP}{TP+FP} \tag{1}$$
$$F1 = \frac{2\times precision \times recall}{precision + recall}$$

However, in the case of unbalanced categories where some categories have a large number of instances, but some categories have much fewer instances, the F1 score does not take into account the difference in the number of instances in different categories. Therefore, to better compare the performance of the overall results, the weighted average F1 value (Equation (2)) is used for a performance indicator. where n represents the number of categories, S_i represents the number of instances of the ith category, T indicates the total number of instances, and $F1_i$ denotes the F1 score of the ith category:

$$weightedF1_{avg} = \sum_{i=1}^{n} \left(\frac{S_i}{T} \times F1_i \right) \tag{2}$$

3. C-BiLSTM-Based Classification Framework

Figure 1 illustrates the framework of the proposed C-BiLSTM-based method for classifying construction accident narratives. There are two modules in the framework, i.e., model training and model application. In the module of model training, labeled training data need first to be pre-processed, such as word segmentation and stop word removing before being transmitted to the C-BiLSTM classifier for training. In the module of the model application, raw data were pre-processed and then transmitted to the trained model, and then corresponding classification labels were obtained to complete the classification task. See Figure 2 for the framework of the C-BiLSTM Model, which mainly comprises two parts, i.e., CNN and BiLSTM. In the model, the convolutional layer extracts n-gram features in the text for sentence modeling. Then BiLSTM obtains the forward and backward context features via the combination of forward and backward LSTMs and transfers the result to the softmax classifier.
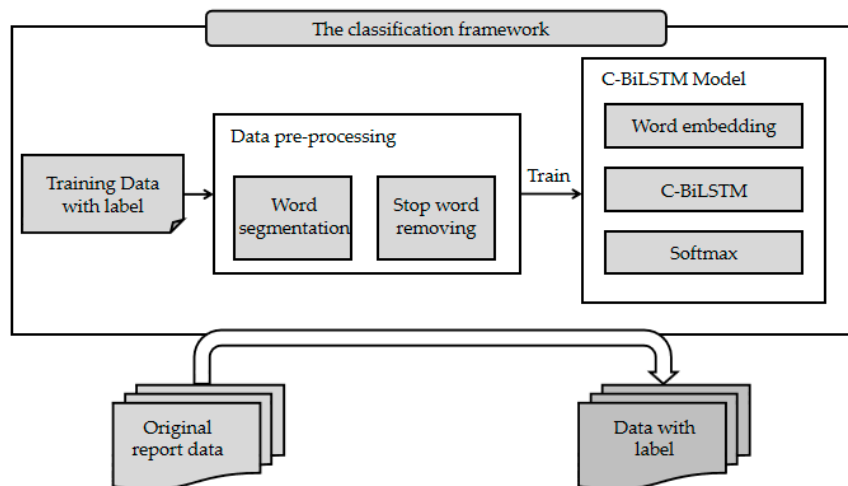
Figure 1. The framework of the convolutional bidirectional long short-term memory (C-BiLSTM)-based classification approach.
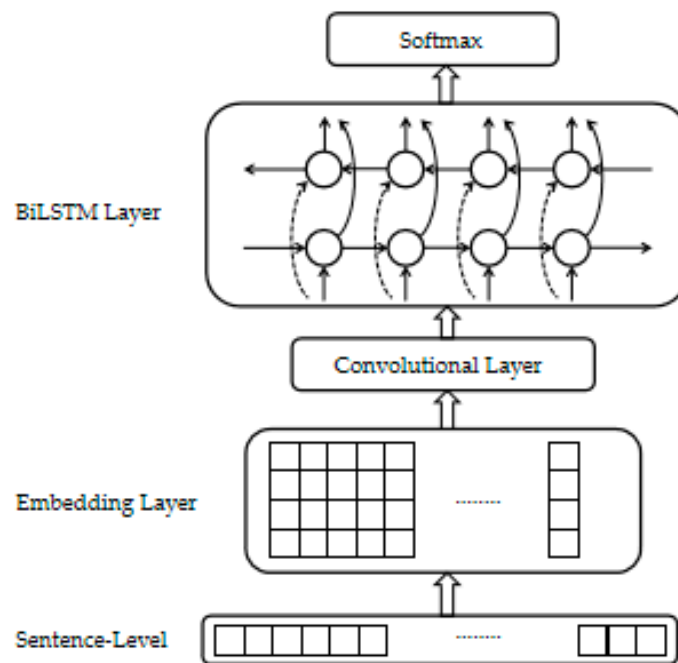
Figure 2. The C-BiLSTM model.

3.1. Data Pre-Processing

In real applications, raw data often contain a lot of noise information which not only affects the accuracy of data mining but also influences the work efficiency. Therefore, a series of pre-processing work needs to be performed on the data before the data are used. The main tasks of text pre-processing are word segmentation and stop word removal. Word segmentation refers to the division of long text data into individual words or phrases. NLTK is a kind of natural language toolkit that is the most commonly used in the NLP field and provides a set of more professional English word segmentation tools [47]. Therefore, in this study, the tokenize segmentation package provided by NLTK was used directly to convert the text data into a word-level dictionary. In-text datasets, the most common words may appear many times, such as "in," "the," and "a," and they do not provide valuable information. Stop word removing means the removal of these words which can effectively reduce the size and dimension of data and will make training faster and better [48]. Other pre-processing works include the conversion of uppercase letters to lowercase letters, removal of numbers and special symbols, and lemmatization. All these can effectively clean the data to improve the accuracy and speed of data mining.

3.2. Word Embedding

Traditional methods of word representation, such as one-hot vectors, usually have two problems which are losing word order and excessive dimension. In this paper, the distributed word vector representation [49] with automatic parameter tuning was employed to replace the one-hot sparse matrix used in traditional machine learning models. It has a better performance in obtaining the semantics and syntactic information of words in each text. The core of the study lies in the text level classification. Suppose a text contains L words and w_i stands for the vector of the ith word in the text, the word representation a_i can be embedded in the matrix via embedded matrix W, which means matrix representation x_i, as shown in Equation (3). Word2vec method proposed by Mikolve et al. [50] is used in this paper for word embedding. The Skip-gram model in the Word2ve method is used for the task, which trains semantic embedding by predicting target words from context and obtains the semantic relation between words.

$$x_i = W{\cdot}a_i \tag{3}$$

Although Word2vec model has achieved good results in many fields, it cannot deal with polysemy well because Word2vec uses unique word vectors to represent multiple semantics of a word. To better deal with the polysemy issue, this study uses the BERT model to pre-train texts. Devlin et al. [34] pre-trained over 3 billion words in BooksCorpus and English Wikipedia using a multi-layer, two-way Transformer encoder, and obtained a BERT pre-training model. To apply the BERT model to text classification of construction accidents, this study directly employed this BERT-base model.

3.3. One Dimension Convolutional Layer

The local features of a text can be extracted via CNN, which refers to a kind of feedforward neural network whose model structure mainly includes an input layer, convolutional layer, pooling layer, fully connected layer, and output layer. In C-BiLSTM Model, a single convolutional layer is used to reduce data dimensions and extract serial information. Its structure is shown in Figure 3. A max-over-pooling or dynamic k-max pooling is generally used to choose the most important or k-most important features after the convolution. However, the input to BiLSTM must be a serialized structure, and pooling will destroy the sequence organization of the text. Therefore, data after convolution operation will not undergo pooling operation anymore.
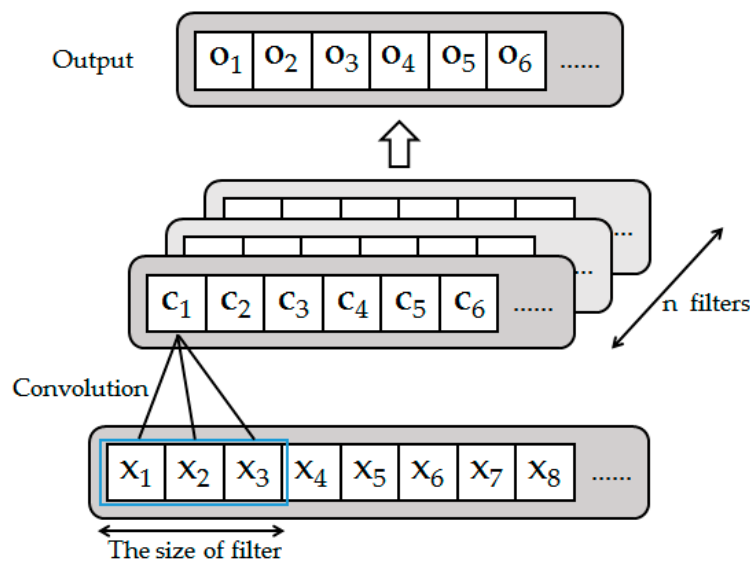

Figure 3. The architecture of the convolution operation.

Take $X \in R^{L \times d}$, the text representation after word embedding, as input and the word vector of each word in the text is $x_i \in R^d$ where L is the max length of the text and d is the dimension of word vector (the dimensionality in this study is 300). Convolution is mainly used for feature extraction. The features of input text can be extracted by sliding on input sequence via filter $m \in R^{k \times d}$ (k is the length of filter). In every position i in the sentence, there is a window vector w_i that contains k continuous word vectors $(x_i, x_{i+1}, \ldots, x_{i+k-1})$. Eigenvalue c_i is obtained by convoluting filter m and window vector w_i and its calculation process is shown in Equation (4).

$$c_i = f(m \cdot w_i + b) \tag{4}$$

where b is bias term, the value of which can be adjusted in the training process; $f(.)$ represents the nonlinear activation function, rectified linear units (ReLU). ReLU has better performance than other activation functions in terms of iterations required for reducing network convergence. By convoluting every window vector in the text, the feature sequence $C = (c_1, c_2, \ldots, c_{L-k+1})$ can be obtained for the

text. 128 filters in the same size are used in C-BiLSTM to acquire several feature sequences. Therefore, data will become a new feature representation O after the convolution layer, as shown below.

$$O = (C_1; C_2; \ldots; C_{128})\tag{5}$$

To be specific, semicolon stands for column vector connection and new features representation will be fed into BiLSTM as input.

3.4. BiLSTM

LSTM model can be used to solve the problem that traditional machine learning models are difficult to extract high-level semantics in texts when classifying texts. This model adopts a text sequence matrix composed of pre-trained distributed word vectors as input and then extracts the feature expressions containing context information by using its unique memory structure. The LSTM model structure is shown in Figure 4a. The standard LSTM network can only leverage the historical context. However, the lack of future context may lead to an incomplete understanding of the meaning of the text. BiLSTM is the combination of a forward LSTM layer and a backward LSTM layer. The information of context can be fully used by summarizing the information of two ideates before and after the word. The model structure is shown in Figure 4b.
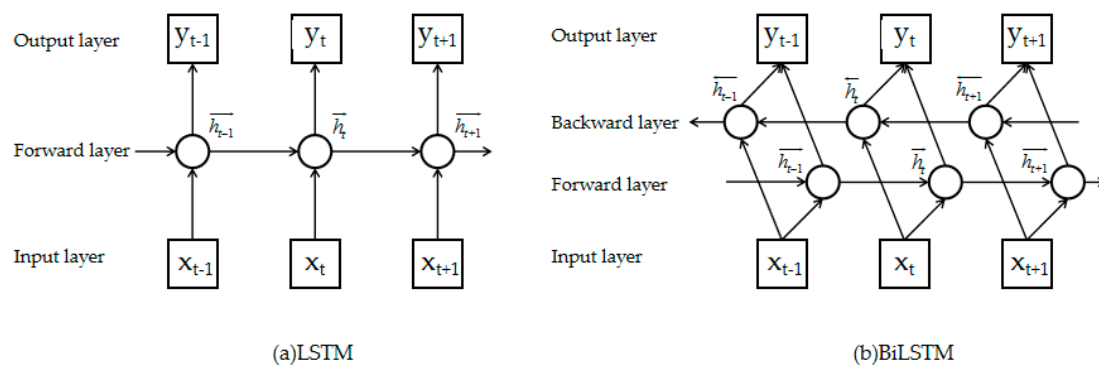

Figure 4. The architecture of the (**a**) LSTM model and (**b**) BiLSTM model.

The key idea of RNN is to use sequential information [51]. Text classification can be treated as a sequential modeling task. Due to its sequential feature, RNN models have been widely used in text classification tasks [52,53]. The LSTM model is a special type of RNN [54] that overcomes the issue of vanishing gradients during RNN model training. The key point is to find and establish long-term dependencies between input values by its specially designed memory unit so that it can understand more contextual information to extract high-level abstract features from texts. The memory unit structure of the BiLSTM model is shown in Figure 5.
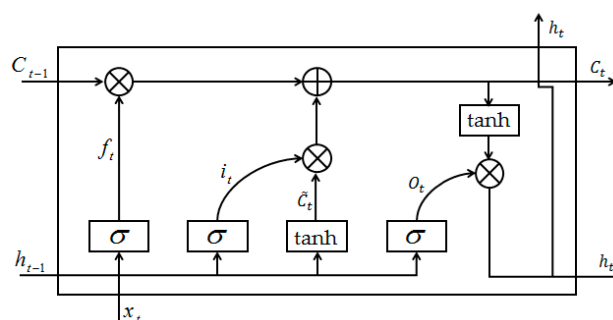

Figure 5. Memory unit structure of BiLSTM.

The most important part of the memory unit is the memory state C which is transmitted directly over the entire structure chain and performs only a small amount of linear operation so that the information can be easily kept unchanged during transmission. At the same time, the memory unit has a smart "gate" structure to add or delete the information contained in the memory state. The so-called "gate" is a method of selecting information, which includes the pointwise multiplication operation of vectors and the sigmoid function. A complete memory unit mainly includes the following parts: memory C_{t-1} at time t−1, output h_{t-1} at time t−1, forget gate f_t, gate i_t, and output gate O_t, among which the values of the three gates are all between 0 and 1, while the memory state C_{t-1} records the historical information of all previous time nodes, which is the long-term memory of the model, and the h_{t-1} records the information of the time node right before the current time, which is the short-term memory of the model. The memory state of the jth memory unit at time t, C_t^j, is the result of the input gate i_t^j, forget gate f_t^j, and the previous memory state C_{t-1}^j. Referring to the version of Zaremba et al. [55], the calculation formula of the memory unit is as follows

$$C_t^j = f_t^j \times C_{t-1}^j + i_t^j \times \widetilde{C}_t^j \tag{6}$$

where

$$f_t = \sigma\left(W_f \cdot [h_{t-1},\ x_t] + b_f\right) i_t = \sigma(W_i \cdot [h_{t-1},\ x_t] + b_i) \widetilde{C}_t = tanh(W_c \cdot [h_{t-1},\ x_t] + b_C) \tag{7}$$

In Equation (7), W represents the weight matrix corresponding to each control gate, b denotes bias parameter, σ denotes sigmoid activation function, tanh represents the hyperbolic tangent function and x_t represents the input of the model at time t. The input gate i_t and the forget gate f_t respectively control the addition of new information and the deletion of old information. When the memory unit is updated, the hidden layer will calculate the current hidden layer h_t according to the result of the current output gate O_t:

$$O_t = \sigma(W_O \cdot [h_{t-1},\ x_t] + b_O) h_t = O_t \times tanh(C_t) \tag{8}$$

From the data processing procedure of the memory unit, it can be known that the core idea of the memory unit structure is to continuously update the long-term and short-term information in the model according to the input information of the current word, that is, to continuously obtain the context features in the text. Let the hidden status output by forward LSTM be $\overrightarrow{h_t}$ and that output by backward LSTM be $\overleftarrow{h_t}$ at moment t, the hidden status output h_t by BiLSTM will be

$$h_t = \overrightarrow{h_t} \oplus \overleftarrow{h_t} \tag{9}$$

The BiLSTM model finally achieves the prediction of the category of the text by using the softmax classification layer, as shown in Figure 2. Softmax refers to the softmax regression model which is a commonly used multi-classification algorithm. It calculates the probability that the text to be classified belongs to each category by transmitting the output of the BiLSTM hidden layer to the softmax classification layer so that the classification result is finally obtained by the maximum probability among all categories.

4. Results and Discussions

4.1. Data Description

The original data of construction accident narratives were downloaded free of charge from the Occupational Safety and Health Administration (OSHA) website which contains more than 16,000 construction accident reports from 1983 to the present. Each report comes with a detailed description of the accident, including the cause and the final result. Unfortunately, these data are not explicitly labeled. Therefore, this study used an open-source dataset published in the early work of Goh et al. [40] which contains 4470 narrative data downloaded from OSHA, where 1000 narratives are labeled. Table 1

shows a labeled example in which the title and the narrative are combined as one single piece of text data to make full use of the header information. The dataset was labeled according to labels used by the Institute for Workplace Safety and Health [56].

Table 1. Example of accident narrative (with label).

Title	Employee is Found Dead after Exposure to Chlorine
Summary	On 27 June 2008 employee #1 and a coworker were performing mold inspections in an army barrack. A contractor was spraying a 6 to 1 mixture of bleach and water. Employee #1 complained of chest pains and the odor of chlorine later that evening. He was found dead in his hotel room the following day.
Label	exposure to chemical substances

By carefully checking the labeled dataset, it is found that there are obvious imbalances in the samples of different categories in the dataset, which tends to results in over-fitting of the category with a large number of samples and the under-fitting of the category with a small number of samples in the process of model training [57]. Therefore, to make the dataset more suitable for classification experiments, this research manually labeled some additional data to make the number of samples in all categories more balanced. The dataset now contains 1863 instances, and it has been published to GitHub [58]. The 11 accident categories and their sample distributions are shown in Table 2.

Table 2. Labels and sample distributions used in this research.

	Label	Description	Count	%
1	Caught In/Between Objects	Fractures, cuts, lacerations, or amputations caused when a worker is caught in between objects, generally referring to hand tools.	95	5%
2	Collapse of Object	Cases resulting from structural failure.	258	14%
3	Electrocution	Direct electric shock or any burns caused by electrical faults.	270	14%
4	Exposure to Chemical Substances	Worker comes into contact with toxic/corrosive chemical substances.	109	6%
5	Exposure to Extreme Temperatures	Extreme temperatures caused by frost, hot liquid, or gases (this category includes hypothermia).	92	5%
6	Falls	Slip or trip cases and cases where a victim falls from a high elevation (not due to structural failure).	293	16%
7	Fires and Explosion	Injuries caused by direct fires and explosion (not including electrical burns).	173	9%
8	Struck by Falling Object	Victim is hit by a falling object (which is not a result of structural failure).	124	7%
9	Struck by Moving Objects	Victim is hit by a moving object (which is not in free fall).	164	9%
10	Traffic	Injury occurs while a worker is driving a vehicle or when a moving vehicle strikes a worker.	169	9%
11	Other	Cases that do not fall in any of the above categories. Some less-frequently occurring categories are merged, as the number of occurrences is very low (drowning, suffocation).	116	6%
	TOTAL		1863	100%

To train the model and evaluate model performance, a separate test dataset needs to be reserved for the evaluation of the universality of the model in analyzing complex texts [59]. Therefore, the labeled dataset was randomly divided into two groups, i.e., one training set for optimizing the model and a test set for evaluating the model performance. The training set contains 1490 instances (accounting for 80% of the total data) and the test set contains 373 instances (accounting for 20% of the total data).

4.2. Baseline Models

In the field of text classification, a number of machine learning algorithms have shown good performance. However, no single algorithm can always be superior to other algorithms and is applicable in all fields [23]. To evaluate the performance of the proposed C-BiLSTM-based method, this study selected three baseline classifiers, namely SVM, NB, and LR, for comparison. These three classifiers are not only widely used, but also state-of-the-art in the field of classification of construction-related documents. These algorithms are well established and are introduced in detail by Bishop [60]. In addition, the results of the C-BiLSTM model were compared with three deep learning models, namely CNN, LSTM, and BiLSTM. To compare the effects of different pre-training models on the experimental results, BERT and Word2Vec were used to process the data as inputs to the C-BiLSTM model.

SVM [61] is established based on the statistics and the principle of risk structure minimization, with the ultimate goal to find the optimal classification line in the current condition, which means the optimal partition hyperplane that achieves optimal classification of the test sample. The optimal partition hyperplane not only correctly divides the subject dataset, but also ensure the partitions reach the maximum interval.

NB [62] has a simple structure and is extensively used. It models the classification of documents through the probability model under the assumption that different items are independent of each other and subject to the same distribution. The algorithm idea of the NB classifier is that, for the given item to be classified, the probability of occurrence of each category under the condition of some given features is solved, and the category of the item to be classified is determined according to the maximum probability.

LR [63], also known as Logistic Regression Analysis, is used to deal with the regression problem where dependent variables are categorical variables. It commonly copes with binary classification or binomial distribution problem but can also deal with multi-classification problems. It predicts the probability of future outcomes through the performance of historical data.

4.3. Experiment Results

As discussed in Section 2, the F1 score is used as an indicator to evaluate the performance of a classifier of construction accident narratives. Table 3 summarizes the F1 results of the proposed C-BiLSTM model and other baseline models, with the highest F1 score (i.e., the best classification performance) underlined. Figure 6 gives a better visualized illustration.
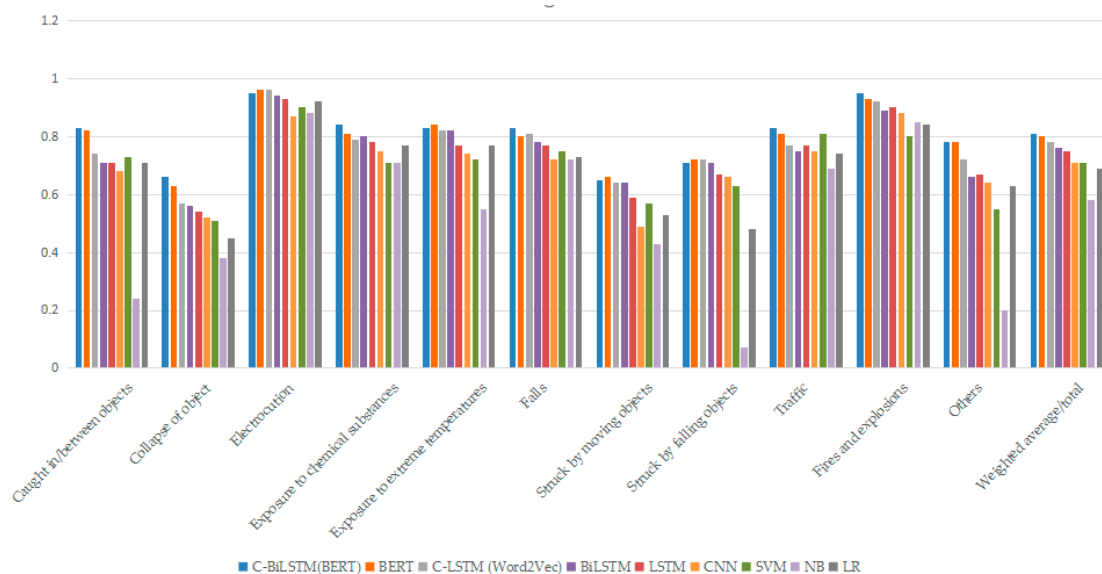

Figure 6. F1 scores of different methods.

Table 3. Comparison of results of the C-BiLSTM-based model and baseline models.

Labels	C-BiLSTM (BERT)	BERT	C-BiLSTM (Word2vec)	BiLSTM	LSTM	CNN	SVM	NB	LR
	F1	F1	F1	F1		F1		F1	F1
Caught in/between objects	0.83	0.82	0.74	0.71	0.71	0.68	0.73	0.24	0.71
Collapse of object	0.66	0.63	0.57	0.56	0.54	0.52	0.51	0.38	0.45
Electrocution	0.96	0.96	0.96	0.94	0.93	0.87	0.90	0.88	0.92
Exposure to chemical substances	0.84	0.81	0.79	0.80	0.78	0.75	0.71	0.71	0.77
Exposure to extreme temperatures	0.83	0.84	0.82	0.82	0.77	0.74	0.72	0.55	0.77
Falls	0.83	0.80	0.81	0.78	0.77	0.72	0.75	0.72	0.73
Struck by moving objects	0.65	0.66	0.64	0.64	0.59	0.49	0.57	0.43	0.53
Struck by falling objects	0.71	0.72	0.72	0.71	0.67	0.66	0.63	0.07	0.48
Traffic	0.83	0.81	0.77	0.75	0.77	0.75	0.81	0.69	0.74
Fires and explosions	0.95	0.93	0.92	0.89	0.90	0.88	0.80	0.85	0.84
Others	0.78	0.78	0.72	0.66	0.67	0.64	0.55	0.20	0.63
Weighted average/total	0.81	0.80	0.78	0.76	0.75	0.71	0.71	0.58	0.69

It can be seen from Table 3 that the proposed C-BiLSTM-BERT-based model is generally superior to other methods in terms of both the F1 score and the weighted average F1 score, with a maximum weighted average F1 of 0.81, which is better than 0.8 of BERT, 0.78 of C-BiLSTM-Word2vec, 0.76 of BiLSTM, 0.75 of LSTM, 0.71 of CNN, 0.71 of SVM, 0.58 of NB, and 0.69 of LR. In addition, the proposed model achieves the highest F1 score for most categories, except for "Exposure to chemical substances", "Struck by moving objects" and "Struck by falling objects" where the BERT method has a better performance respectively. It is worth noting that the category "Electrocution" shows the best classification results in all methods, especially in the C-BiLSTM-BERT-based model with a superior score of 0.96, and the category "Struck by moving objects" has the worst classification result, with an F1 score of 0.65 in the proposed method. In general, different classifiers can have various performances in the classification results of different categories, and the proposed C-BiLSTM-based method is better in an aggregate performance.

4.4. Discussions and Future Work

It is obvious that the C-BiLSTM-based method outweighs a single model in terms of the classification effect. As the convolutional layer can capture the local correlation of spatial or temporal structures, it can extract n-gram features of different positions of text from word vectors by the convolutional filter. What is more, the BiLSTM network can capture the features of a long-time interval in the preceding part and obtain the features of the following text. Therefore, BiLSTM has more advantages than LSTM in making full use of semantic and word order information of the text. The C-BiLSTM-based method has the strength of both models to improve classification accuracy. It is worth noting that compared with the SVM model, a single CNN does not significantly improve the text classification effect, while the BiLSTM model improves the accuracy of classification results more. For the text classification task, the word order feature combined with context may have more influence on the classification result than the local feature. It is noted that the results of the BERT model are much better than those of the C-BiLSTM model, which uses Word2vec as the data pretraining. The results

of the C-BILSTM model using BERT as a data pretraining method have also been clearly improved. The results indicate that the generalization ability of BERT and the extraction of context features have obvious advantages, but the complexity of the model and the time of data processing are much higher than the Word2vec model.

It can be found that in Table 3, although the overall performance of this proposed C-BiLSTM model is better than other baseline methods, the results are not ideal in the classification of some specific categories, which greatly affects the weighted F1 score. One example is the category of "Collapse of object" which only gets an F1 score of 0.57. One possible reason is the unclear definition of the classification label. By carefully examining the dataset, it was found that some labels are even difficult to be classified by human readers, especially the three types: "Collapse of objects," "Falls," and "Struck by falling objects". In accident narratives, the accidents involving the collapse of objects are often accompanied by accidents in which workers fall and are hit by falling objects. The inaccuracy of the definition of this classification label may affect the accuracy of model identification. On the other hand, we noticed that the accident narratives often contain too much description of the working environment and the contents of rescue and human care after the accidents, while there are not enough words describing the direct cause of the accident. These redundant texts also pose a challenge for the accuracy of the classification result.

It can be found from the experimental results that no classifier can achieve the consistent and best performance for all categories. As such, in future work, the performance of different DL models can be further investigated. In addition, a common problem in text classification tasks is too many terminologies in texts. To reduce the number of features and improve performance, professional ontology and dictionaries can be used in the pre-processing [39] to help remove the detailed description which is not related to events and reduce data dimensions. This work usually requires experts to build an ontology and/or a dictionary of a specific domain.

Finally, this proposed C-BiLSTM-based model was only used on the text dataset obtained from OSHA for evaluation. As such, the consistency of models in different data sets can be further discussed in future work.

5. Conclusions

Accidents could occur every day on a construction site and cannot be ignored by construction industry practitioners. To make full use of the effective information and knowledge in existing construction accident reports to prevent future similar accidents, this paper proposed an automatic classification method for the causes of accidents, based on a C-BiLSTM model. By extracting the semantic features from the contextual information in the text, this method can automatically classify the construction accident narratives according to its causes. This proposed method shows significant improvement in terms of classification performance measured by the F1 score, compared to other baseline methods in the construction industry such as SVM, NB, and LR classifiers. More accurate classification results obtained by this C-BiLSTM-based method also provides a data basis for the further application of the construction accident reports, such as in the prediction of construction accidents [64]. This classification model also provides a reference for the text classification in other specific fields in the construction industry, for example, quality inspection.

In addition, this work still requires manual labeling of construction accidents to form datasets for model training. In future work, unsupervised models that do not need to be labeled can be considered for experiment and comparison of results. On the other hand, through the analysis of construction reports, it can be found that real accidents are often caused by multiple accident causes. Therefore, the processing of data according to multi-label classification tasks or further refinement of classification categories can be considered in future research.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Al-Humaidi, H.M.; Tan, F.H. Construction safety in Kuwait. *J. Perform. Constr. Facil.* **2010**, *24*, 70–77. [CrossRef]

2. International Labor Organization (ILO). Available online: http://www.ilo.org/global/topics/safety-and-health-at-work/lang--en/index.html (accessed on 2 June 2019).

3. Sacks, R.; Rozenfeld, O.; Rosenfeld, Y. Spatial and temporal exposure to safety hazards in construction. *J. Constr. Eng. Manag.* **2009**, *135*, 726–736. [CrossRef]

4. Abdelhamid, T.S.; Everett, J.G. Identifying root causes of construction accidents. *J. Constr. Eng. Manag.* **2000**, *126*, 52–60. [CrossRef]

5. Zhou, Z.; Yang, M.G.; Li, Q. Overview and analysis of safety management studies in the construction industry. *Saf. Sci.* **2015**, *72*, 337–350. [CrossRef]

6. Hallowell, M.; Gambatese, J. A formal model for construction safety risk management. In Proceedings of the Construction and Building Research Conference of the Royal Institution of Chartered Surveyors, COBRA 2007, Atlanta, GA, USA, 10–14 September 2007.

7. Esmaeili, B.; Hallowell, M.R. Diffusion of Safety innovations in the construction industry. *J. Constr. Eng. Manag.* **2012**, *138*, 955–963. [CrossRef]

8. Teizer, J.; Allread, B.S.; Fullerton, C.E.; Hinze, J. Autonomous pro-active real-time construction worker and equipment operator proximity safety alert system. *Autom. Constr.* **2010**, *19*, 630–640. [CrossRef]

9. Baradan, S.; Usmen, M.A. Comparative injury and fatality risk analysis of building trades. *J. Constr. Eng. Manag.* **2006**, *132*, 533–539. [CrossRef]

10. Hallowell, M.R.; Gambatese, J.A. Activity-based safety risk quantification for concrete formwork construction. *J. Constr. Eng. Manag.* **2009**, *135*, 990–998. [CrossRef]

11. Shapira, A.; Lyachin, B. Identification and Analysis of Factors Affecting Safety on Construction Sites with Tower Cranes. *J. Constr. Eng. Manag.* **2009**, *135*, 24–33. [CrossRef]

12. Esmaeili, B.; Hallowell, M.R. Using network analysis to model fall hazards on construction projects. *Saf. Health Constr.* **2011**, *99*, 24–26.

13. Desvignes, M. Requisite Emperical Risk Data for Integration of Safety with Advanced Technologies and Intelligent Systems. Master's Thesis, Department of Civil, Environmental, and Architectural Engineering, University of Colorado, Boulder, CO, USA, 2014.

14. Chua, D.K.H.; Goh, Y.M. Incident causation model for improving feedback of safety knowledge. *J. Constr. Eng. Manag.* **2004**, *130*, 542–551. [CrossRef]

15. Aggarwal, C.C.; Zhai, C.X. A Survey of Text Classification Algorithms. In *Mining Text Data*; Aggarwal, C., Zhai, C., Eds.; Springer: Boston, MA, USA, 2012; pp. 163–222. ISBN 978-1-4614-3222-7.

16. Zhang, F.; Fleyeh, H.; Wang, X.; Lu, M. Construction site accident analysis using text mining and natural language processing techniques. *Autom. Constr.* **2019**, *99*, 238–248. [CrossRef]

17. Allahyari, M.; Pouriyeh, S.; Assefi, M.; Safaei, S.; Trippe, E.D.; Gutierrez, J.B.; Kochut, K. A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques. *arXiv* **2017**, arXiv:1707.02919.

18. Gantz, J.; Reinsel, D. The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East. Available online: http://www.emc.com/leadership/digital-universe/2012iview.index.htm (accessed on 21 May 2020).

19. Williams, T.P.; Gong, J. Predicting construction cost overruns using text mining, numerical data and ensemble classifiers. *Autom. Constr.* **2014**, *43*, 23–29. [CrossRef]

20. Wang, S.; Manning, C.D. Baselines and bigrams: Simple, good sentiment and topic classification. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, Jeju, Korea, 8–14 July 2012; Volume 2, pp. 90–94.

21. Watanabe, A.; Sasano, R.; Takamura, H.; Okumura, M. Generating personalized snippets for web page recommender systems. In Proceedings of the 2014 IEEE/WIC/ACM International Conference on Web Intelligence, Intelligence Agent Technolog Work, WI-IAT 2014, Warsaw, Poland, 11–14 August 2014; Volume 2, pp. 198–208.

22. Schwartz, H.A.; Eichstaedt, J.C.; Kern, M.L.; Dziurzynski, L.; Ramones, S.M.; Agrawal, M.; Shah, A.; Kosinski, M.; Stillwell, D.; Seligman, M.E.P.; et al. Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. *PLoS ONE* **2013**, *8*, 1631–1642. [CrossRef]

23. Sebastiani, F. Machine Learning in Automated Text Categorization. *ACM Comput. Surv.* **2002**, *34*, 1–47. [CrossRef]

24. Kecman, V. *Support Vector Machines–An Introduction*; Springer: Berlin/Heidelberg, Germany, 2005; pp. 1–47.

25. Russell, S.J.; Norvig, P. *Artificial Intelligence: A Modern Approach*, 2nd ed.; Prentice Hall: Englewood Cliffs, NJ, USA, 2003; pp. 1–1095.

26. Dasarathy, B. *V Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques*; IEEE Computer Society Press: Los Alamitos, CA, USA, 1991; pp. 217–224. ISBN 9780818659300.

27. Deng, L.; Yu, D. Deep Learning: Methods and Applications. *Found. Trends Signal. Process.* **2014**, *7*, 197–387. [CrossRef]

28. Hinton, G.E.; Salakhutdinov, R.R. Reducing the Dimensionality of Data with Neural Networks. *Science* **2006**, *313*, 504–507. [CrossRef]

29. Zhong, B.; Xing, X.; Love, P.; Wang, X.; Luo, H. Convolutional neural network: Deep learning-based classification of building quality problems. *Adv. Eng. Inform.* **2019**, *40*, 46–57. [CrossRef]

30. Liu, G.; Guo, J. Bidirectional LSTM with attention mechanism and convolutional layer for text classification. *Neurocomputing* **2019**, *337*, 325–338. [CrossRef]

31. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, L. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.

32. Lu, J.; Yang, J.; Batra, D.; Parikh, D. Hierarchical question-image co-attention for visual question answering. *Proc. Adv. Neural Inf. Process. Syst.* **2016**, *29*, 289–297.

33. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 3146–3154.

34. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pretraining of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.

35. Bertke, S.J.; Meyers, A.R.; Wurzelbacher, S.J.; Bell, J.; Lampl, M.L.; Robins, D. Development and evaluation of a Naïve Bayesian model for coding causation of workers' compensation claims. *J. Saf. Res.* **2012**, *43*, 327–332. [CrossRef] [PubMed]

36. Tanguy, L.; Tulechki, N.; Urieli, A.; Hermann, E.; Raynal, C. Natural language processing for aviation safety reports: From classification to interactive analysis. *Comput. Ind.* **2016**, *78*, 80–95. [CrossRef]

37. Marucci-Wellman, H.R.; Corns, H.L.; Lehto, M.R. Classifying injury narratives of large administrative databases for surveillance—A practical approach combining machine learning ensembles and human review. *Accid. Anal. Prev.* **2017**, *98*, 359–371. [CrossRef]

38. Abdat, F.; Leclercq, S.; Cuny, X.; Tissot, C. Extracting recurrent scenarios from narrative texts using a Bayesian network: Application to serious occupational accidents with movement disturbance. *Accid. Anal. Prev.* **2014**, *70*, 155–166. [CrossRef]

39. Tixier, A.J.P.; Hallowell, M.R.; Rajagopalan, B.; Bowman, D. Automated content analysis for construction safety: A natural language processing system to extract precursors and outcomes from unstructured injury reports. *Autom. Constr.* **2016**, *62*, 45–56. [CrossRef]

40. Goh, Y.M.; Ubeynarayana, C.U. Construction accident narrative classification: An evaluation of text mining techniques. *Accid. Anal. Prev.* **2017**, *108*, 122–130. [CrossRef]

41. Lu, C.; Huang, H.; Jian, P.; Wang, D.; Guo, Y. A P-LSTM neural network for sentiment classification. In Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining, PAKDD, Jeju Island, Korea, 23–26 May 2017; pp. 524–533.

42. Wang, Y.; Feng, S.; Wang, D.; Zhang, Y.; Yu, G. Context-aware chinese microblog sentiment classification with bidirectional LSTM. In Proceedings of the Asia-Pacific Web Conference on Web Technologies and Applications, APWeb, Suzhou, China, 23–25 September 2016.

43. Wei, X. A convolution-LSTM-based deep neural network for cross-domain MOOC forum post classification. *Information* **2017**, *8*, 92. [CrossRef]

44. Le, T.; Bui, G.; Duan, Y. A multi-view recurrent neural network for 3D mesh segmentation. *Comput. Graph.* **2017**, *66*, 103–112. [CrossRef]

45. Yenala, H.; Chinnakotla, M.; Goyal, J. Convolutional Bi-directional LSTM for Detecting Inappropriate Query Suggestions in Web Search. In Proceedings of the Advances in Knowledge Discovery and Data Mining, PAKDD, Jeju Island, Korea, 23–26 May 2017; pp. 3–16.

46. Buckland, M.; Gey, F. The relationship between Recall and Precision. *J. Am. Soc. Inf. Sci.* **1994**, *45*, 12–19. [CrossRef]

47. Bird, S.; Loper, E. NLTK: The Natural Language Toolkit. In Proceedings of the ACL Interactive Poster & Demonstration Sessions, Barcelona, Spain, 21–16 July 2004.

48. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.; Dean, J. Distributed representations ofwords and phrases and their compositionality. *Adv. Neural Inf. Process. Syst.* **2013**, *26*, 1–9.

49. Wang, P.; Xu, B.; Xu, J.; Tian, G.; Liu, C.L.; Hao, H. Semantic expansion using word embedding clustering and convolutional neural network for improving short text classification. *Neurocomputing* **2016**, *174*, 806–814. [CrossRef]

50. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient Estimation of Word Representations in Vector Space. *arXiv* **2013**, arXiv:1301.3781.

51. Funahashi, K.I.; Nakamura, Y. Approximation of dynamical systems by continuous time recurrent neural networks. *Neural Netw.* **1993**, *6*, 801–806. [CrossRef]

52. Wang, L.; Wang, Z.; Liu, S. An effective multivariate time series classification approach using echo state network and adaptive differential evolution algorithm. *Expert Syst. Appl.* **2016**, *43*, 237–249. [CrossRef]

53. Cao, W.; Song, A.; Hu, J. Stacked residual recurrent neural network with word weight for text classification. *IAENG Int. J. Comput. Sci.* **2017**, *44*, 277–284.

54. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]

55. Zaremba, W.; Sutskever, I.; Vinyals, O. Recurrent Neural Network Regularization. *arXiv* **2014**, arXiv:1409.2329.

56. Workplace Safety and Health Institute. Available online: https://www.wsh-institute.sg/ (accessed on 3 June 2019).

57. He, H.; Garcia, E.A. Learning from imbalanced data. *IEEE Trans. Knowl. Data Eng.* **2009**, *21*, 1263–1284.

58. The OSHA Dataset Used in This Article. Available online: https://github.com/LemonDa/OSHA_Dataset (accessed on 3 June 2019).

59. Saptoro, A.; Tadé, M.O.; Vuthaluru, H. A modified Kennard-Stone algorithm for optimal division of data for developing artificial neural network models. *Chem. Prod. Process. Model.* **2012**, 7. [CrossRef]

60. Bishop, C.M. *Pattern Recognition and Machine Learning (Information Science and Statistics)*; Springer: Berlin/Heidelberg, Germany, 2006; pp. 1–55.

61. Liu, Y.; Bi, J.W.; Fan, Z.P. A method for multi-class sentiment classification based on an improved one-vs-one (OVO) strategy and the support vector machine (SVM) algorithm. *Inf. Sci.* **2017**, *394*, 38–52. [CrossRef]

62. Kononenko, I. Semi-naive bayesian classifier. *Lect. Notes Comput. Sci.* **1991**, *482*, 206–219.

63. Scott, A.J.; Hosmer, D.W.; Lemeshow, S. *Applied Logistic Regression*; Wiley: New York, NY, USA, 2000; pp. 1–68.

64. Tixier, A.J.P.; Hallowell, M.R.; Rajagopalan, B.; Bowman, D. Application of machine learning to construction injury prediction. *Autom. Constr.* **2016**, *69*, 102–114. [CrossRef]