Expert Systems With Applications 210 (2022) 118352

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Introducing attentive neural networks into unconventional oil and gas violation analysis and emergency response system

Dan Bi [a,1,*], Ju-e Guo [b]

[a] School of Management, Xi'an Jiaotong University, Xi'an 710049, China
[b] School of Management, Xi'an Jiaotong University, No.28, Xianning West Road, Beilin District, Xi'an 710049, China

ARTICLE INFO

Keywords:
Unconventional oil and gas
Violation analysis
Attentive neural network
Emergency response
Decision support system

ABSTRACT

With the prosperous development of unconventional oil and gas (UOG) began in the mid-1990 s, the proliferation of digital textual compliance reports from the UOG production life-cycle makes it imperative for experts to develop efficient ways of supporting emergency responses based on the textual based data sources. In this respect, we utilized the UOG compliance reports from the Pennsylvania Department of Environmental Protection from 2000 to 2019, then established an attentive neural-network framework to support on-site emergency responses. The advantages of attentive-based neural networks over the other mechanisms are that it not only generates powerful contextual vectors for follow-up tasks but also it allows us to observe the importance of violation factors with respect to different scenarios. The experimental results show that our model can extract valid representation from narrative texts in UOG violation compliance reports and achieve high performance in emergency response. At the same time, we obtained two intriguing practical implications: first, geographical and time characteristics are powerful indicators for supporting decision making in UOG on-site emergency responses; second, there is an urgent need for governments to implement different inspection strategies according to unique UOG sites rather than counties concerning specific geological features, which benefits from saving human labor and financial expenditures.

1. Introduction

Unconventional oil and gas (UOG) developed in the mid-1990 s. After five years, it has been recognized as a "game-changer" for the U.S. natural gas market. However, compared with the masses' enthusiasm for its unprecedented development trend, the in-depth reflection on its environmental threats is far from enough. The environmental hazards caused during its production stage include water shortage, spills of chemicals, contamination of surface/groundwater, earthquake, and soil erosion (Torres et al., 2016). Countries have built the UOG compliance reports dataset that records specific circumstances of accidents to learn the regular pattern with respect to kinds of environmental hazards. These resources enable operation experts, academics, and a wide range of professionals to conduct in-depth investigations in safety management, environmental assessment, and accident prevention initiatives (Abualfaraj et al., 2016; Guo et al., 2019; Kim & Oliver, 2017; Manda et al., 2014; McKenzie et al., 2010).

The governments have established the UOG violations dataset, which is composed of compliance reports that described the overall on-site condition, cause, fault, source, and enforcement (D. Bi et al., 2021), etc. The recording process is shown in Fig. 1: First, the inspector patrolled the site on regular inspection duty, and recorded the weather, temperature, wind direction, and on-site activities into narrative comments in the "inspection comment". When the inspector noticed a violation, the specific narrative regulation describing the on-site situation was recorded as a "violation comment". After that, the regional government issued corresponding administrative orders (recorded in the "enforcement code") to the site according to the impact and severity of this violation.

Many analysis frameworks have sprung up in the literature for extracting the cause, sources, and pollution mediums from historical compliance reports (Abualfaraj et al., 2016), which lay an essential foundation for accident analysis and supervision systems. Extracting features from narrative compliance reports into new dummy variables in a manual manner has become an increased interest in empirical studies recently, such research objects like environmental regulation on

* Corresponding author at: School of Management, Xi'an Jiaotong University, No.28, Xianning West Road, Beilin District, Xi'an 710049, China.
E-mail addresses: bidan19941126@stu.xjtu.edu.cn (D. Bi), guojue@xjtu.edu.cn (J.-e. Guo).
[1] 0000-0002-5034-5440.

https://doi.org/10.1016/j.eswa.2022.118352
Received 8 February 2022; Received in revised form 23 July 2022; Accepted 31 July 2022
Available online 11 August 2022
0957-4174/

punishment (Guo et al., 2019), environmental regulation, and firm behaviors (Kim & Oliver, 2017), the evolution of multi-well pad development and influence (Manda et al., 2014), violation rates and development decisions (Rahm et al., 2015), etc. Specifically, the fine-grained information can be extracted from narratives that coded data cannot satisfy (Abdat et al., 2014; Ali et al., 2021; Suh, 2021). It leaves the missing gap, that how to extract more useful information from UOG compliance reports in an efficient manner, needs to be filled.

Moreover, the actual inspection pattern is restricted strictly by the workforce and material resources. The existing regular supervision pattern cannot cover the occurrence of emergencies completely while dispatching inspectors to perform supervision is time-consuming and labor-intensive. The deficiencies of existed inspection pattern motivate us to build an end-to-end neural network for supporting UOG on-site emergency responses. While there are two critical challenges to achieving our goals:

Firstly, manually extracting information from narratives is too time-consuming to fit in massive data nowadays, even if it is one of the most accurate methods in feature engineering. In recent years, neural networks like LSTM (Long-short term memory), RNN (Recurrent neural network), and GRU (Gated recurrent unit) can learn the representations that contain semantic and sentiment features in narratives. Specifically, applying neural networks for feature extraction and representation learning from narratives is widely used in driver drossiness analysis (Jacobé de Naurois et al., 2019), and travel recommendation systems (Zhu et al., 2021), and traffic accident forecasting (Ali et al., 2021), etc. Hence, building an efficient and concise framework for supporting the UOG on-site emergency responses is the first challenge in this paper.

Secondly, the decision support system should face the life-cycle of violation orders. In the life-cycle of violation orders, there are three phases apparently as shown in Fig. 1, that is, the inspection of the on-site situation (Identification), the evaluation of violation severity (Evaluation), and the administrative enforcement of violation (Treatment). The information obtained and the measurement executed are different in each stage. Hence, independent neural networks need to be trained and formed unique representations from narratives to improve the performance and efficiency of the decision-making system.

To address the above two challenges, we proposed attentive neural networks to enhance the ability and efficiency of UOG on-site emergency responses. It contains three independent neural networks in the following stages: the identification stage, the evaluation stage, and finally the treatment stage. Specifically, each stage remains a similar neural network with three crucial components: The Bi-GRU (Bidirectional Gated Recurrent Unit) is used to learn the document-level representations from narratives, the attention mechanism used to evaluate the weights of inputs and generate the contextual vectors, and finally the multiple layers perceptions (MLP) used to classification. It is also worth noting that the real-life compliance reports dataset used in this paper has many potentials to help the government unmask hidden patterns in hazards, and eventually establish a decision support system facing the upcoming challenge in the future, where millions, if not billions of records will be generated from online public supervision.

2. Related literature

UOG compliance reports are essential to discern complex patterns in violation analysis. However, extracting key information from narratives and coded data is challenging in this field. In addition, the framework that can reveal the significance of different factors and provide valuable advice in decision-making is another challenging task for academics and professionals. Hence, we will first introduce the violation analysis based on compliance reports, then the neural networks-based and machine learning-related applications for supporting decision making are illustrated.

2.1. Unconventional oil and gas violations analysis based on compliance reports

Hydraulic fracturing and directional and horizontal well drilling during UOG development is the primary source of environmental and human health accidents and threats (Deziel et al., 2020). The known damage includes soil erosion, groundwater pollution, and geological damage (Torres et al., 2016). Hence, it is needed to conduct in-depth investigations to clarify the complex factors in various accidents based on the compliance reports launched by regional environment protection agencies (Abualfaraj et al., 2016; Kim & Oliver, 2017; Rahm et al., 2015).

In recent years, researchers have utilized methods to extract essential information to explore the complex correlations in accidents from compliance reports, which formed a solid foundation for further research in UOG regulation and decision support systems (Dan Bi et al., 2021). Olmstead et al., (2013) conducted a large-scale examination to evaluate the activities that affect the surface water quality during the UOG production in Pennsylvania. The results supported establishing efficient wastewater management plants for governments and practitioners. Moreover, Manda et al., (2014) pay attention to the hypothesis

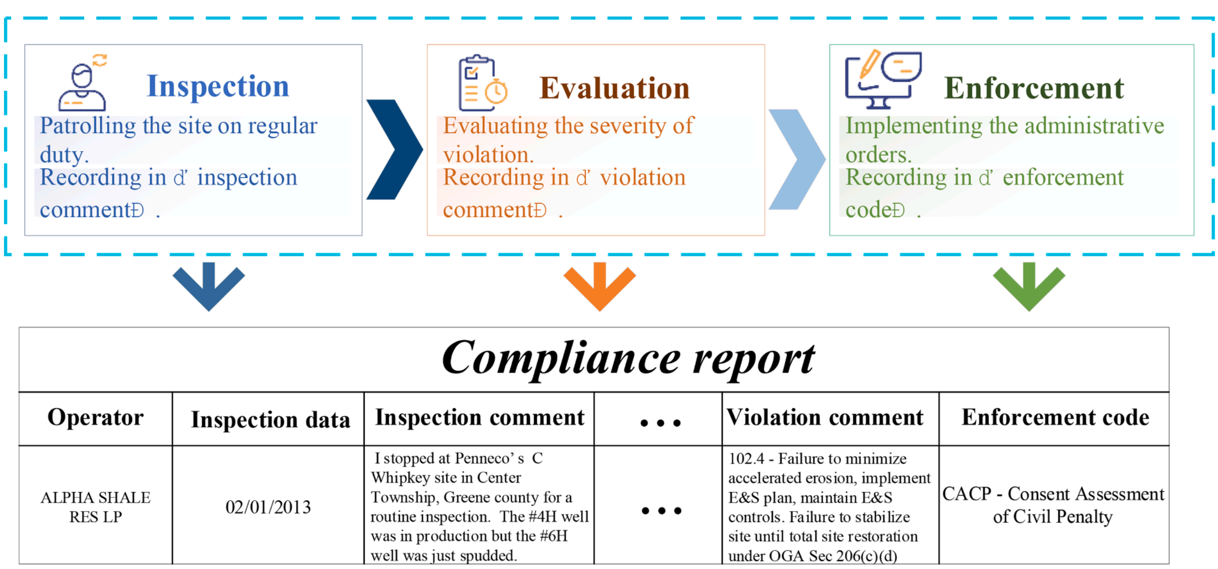

Fig. 1. The process of recording violations and partial compliance reports.

that multi-well pads development can cause fewer environmental violations, then utilized geospatial techniques and statistical analysis based on compliance reports in Pennsylvania to testify to this hypothesis. In recent years, the UOG environmental violation has been the front-burner issue for academics and practitioners. In conclusion, there are several hot research tendencies, that is, water resources-related violations (Rahm et al., 2015), the relationship between monetary provisions and violations (Kim & Oliver, 2017), and on-site storage-related violations (Kuwayama et al., 2017), etc. It is imperative to integrate various violations and then construct a solid decision support system, both for governments and practitioners.

2.2. Application of neural networks in decision support systems

Methodological frameworks with respect to neural networks and machine learning methods are widely used in accident prediction and prevention. The main reason is that the compliance reports contain narratives that are hard to label manually. Moreover, sensor-based systems provide limited information and may fall due to long detection times and high false-alarm rates (Ali et al., 2021). Hence, it is urgent to extract helpful information from narratives (whether from compliance reports or online user-generated comments) and establish a decision support system that can respond to and prevent emergencies on time.

In recent years, a large amount of violation analysis research based on semi-narrative appeared along with the rapid development of machine learning (ML), which mainly focus on workplace safety and health accident (Goh & Ubeynarayana, 2017), road and vehicle safety (Jacobé de Naurois et al., 2019; Roque et al., 2019; Ryder et al., 2017; Ye et al., 2017), automobile insurance fraud (Wang & Xu, 2018), etc. Machine learning has natural advantages in handling unstructured data, it also has been widely used in other domains. Occhipinti et al., (2022) designed a pipeline and evaluate 12 machine learning methods of text classification on the Enron dataset, which greatly benefits experts to carry out subsequent technical innovations in this domain. Esfahlani et al., (2022) adopted Support Vector Machines (SVM) and K-Nearest Neighbors (KNN) in clinical decision-making problems, the results demonstrated that KNN has great potential with respect to predicting the subject's functional improvements. It is obvious that ML methods, which own high performances and interpretable results on regressive and classification tasks, have a promising future for the widespread applications in decision support systems.

There had been an increased interest in automatic classification or auto-coding of accident narratives through the application of text mining and neural networks (Goh & Ubeynarayana, 2017). Ali et al., (2021) applied ontology and latent Dirichlet allocation (OLDA) and bidirectional long short-term memory (Bi-LSTM) to traffic accident prediction through online narratives data, demonstrating that the neural network-based framework is more efficient than other existing systems. Kwayu et al., (2021) utilized structural topic modeling (STM) and network topology to generate and examine the prevalence and interaction of themes from crash narratives. This framework reached high accuracy, emphasizing the significance of both machine learning methods and narrative texts. To handle the dilemma of patronage that long and uncertain waiting time in bus systems, Ma et al., (2022) applied the Multi-attention Graph Neural Networks with long short-term memory (LSTM) for city-wide bus travel time estimation, which addressed the challenge of sparse records and limited data. Jia et al., (2022) constructed an attention-based convolutional network for restaurants to recommend cuisine to customers. The attention mechanism has been widely used in various domains because of its impressive interpretability and excellent performance on downstream tasks. However, as far as we know, no research introduced such an end-to-end attention-based neural network into the UOG violation supervision domain. That is to say, this field is suffering the loss of precious data, it is imperative to build a decision support system which can fully meet the needs of efficient emergency responses and post-accident analysis.

3. The attention-based neural network

In this section we first introduce the attention mechanisms, which have become an integral part of compelling sequence modeling and transduction models in various tasks, allowing modeling of dependencies without regard to their distance in the input or output sequences (Vaswani et al., 2017). It was illustrated concisely by Alammar (2018), and we use this mechanism to discern the key factors in different violation stages. The example and calculation process of the attention mechanism (partially adapted from Alammar J) in our work is presented in Fig. 2: the right side shows the importance of factors in a specific scene calculated by the attention mechanism. The attention weight indicates the degree of the influence of the factor on the dependent variable in the target stage. In this example, the attention weight of 'site' is 0.48, which accounts for the highest proportion of all elements, that implicates 'site' is the most important factor in this circumstance; while the left side takes 'site' and 'year' as value tensor to show the calculation process of attention mechanism, where W^Q and W^V are learnable parameterized matrixes, see more details in Bahdanau et al., (2014) and section 3.3 to 3.5.

3.1. Problem formulation

The theoretical framework proposed in this paper is based on the actual process of violation orders, forming three-layer independent neural networks as shown in Fig. 3: in the first layer, namely the identification phase, it analyzes the narratives reported by inspectors and the public to confirm whether there is a violation occurred at the scene; in the second layer, namely the evaluation phase, it receives the narratives in inspection and violation comments to evaluate the severity of the on-site violation; in the third layer, namely the treatment phase, it gives the suggestions on administrative enforcement according to the specific regulation and violation circumstances.

Hence, our proposed framework contains three independent neural networks, as shown in Fig. 4, which are constructed according to the concept of Fig. 3, the reading order of Fig. 4 is bottom-up, left-to-right. The structure of each stage is similar, hence we take the first stage, namely the identification phase, as an example to explain the information passing flow: 1) first, there are five inputs in this stage, that is, the 'operator', the 'inspection comment', the 'year', the 'site', and the 'county'. The 'inspection result' is the output, which is a binary variable; 2) second, the 'inspection comment' is a variable composed of narrative texts, we first turn word-IDs into word embeddings then pass through two fully connected layers and Bi-GRU to form the final 'inspection comment' document-level embeddings$h_{inspection}$; 3) the 'year', 'site', 'county', and 'operator' are categorical variables, they first turn to category embeddings, then pass through two fully connected layers to form the final categorical embeddings x_n^y, x_n^s, x_n^c andx_n^o, more details are in section 3.2; 4) attention mechanism is applied to form the context vector; 5) the last layer is the fully connected layer, where we applied sigmoid function in this layer. The output is the inspection result, which is a binary variable, that contains 'No violation' and 'Notice of violation' two results.

There are several commonly used components among all phases, we illustrate them in this section for brevity.

Inspection encoder:

The input of this encoder is the sentence describing inspection situation consisting of k words$S_{inspection} = \{w_1, w_2, ..., w_k\}$. Four layers learn the representation, the first three layers represent word-level and the last layer represents document-level. With an embedding matrixW_w^I, the sentence is converted to a sequence of vectors$E^I = \{e_1^I, e_2^I, ..., e_k^I\}$, where, $W_w^I \in R^{d\times|V^I|}$, d is the dimension of the word embeddings and $|V^I|$ is the vocabulary size of inspection. The second and third layers are multilayer

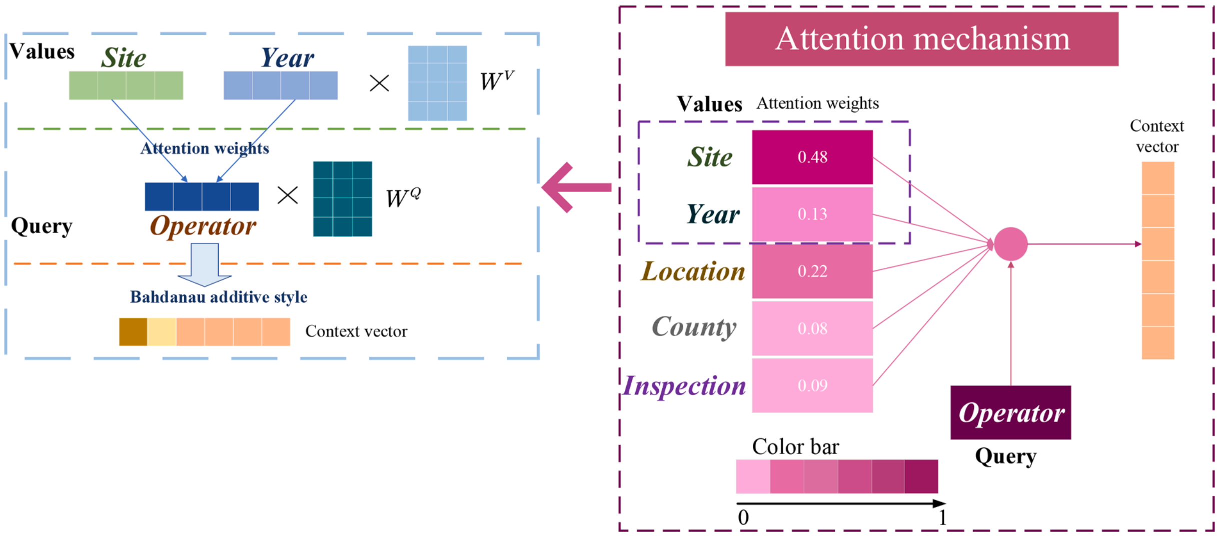

Fig. 2. An illustration of the attention mechanism in our study.

perception (MLP) which is used to learn the hidden representation of the inspection situation:

$$x_n^i = ReLu(W_2^I \cdot ReLu(W_1^I e_n^I + b_1^I) + b_2^I)$$

where $W_2^I, W_1^I, b_1^I, b_2^I$ are weight matrices and bias vectors in the inspection encoder, x_n^i denoted the n^{th} hidden representation of the word w_n in inspection comments.

The fourth layer is a Bidirectional Gate Recurrent Unit (Bi-GRU) network. It is used to learn document-level representation. This Bi-LSTM layer takes the hidden word embeddings as input and outputs hidden states. Let $\overrightarrow{h}_n^t$ and $\overleftarrow{h}_n^t$ denote the t^{th} forward and backward hidden layer representation of the wordw_n, which are calculated by:

$$\overrightarrow{h}_n^t = \overrightarrow{GRU}(x_n^i), \overleftarrow{h}_n^t = \overleftarrow{G}RU(x_n^i)$$

the final hidden layer vector of the wordw_k, denoted as$h_{inspection} = \left[\overrightarrow{h}_i^t; \overleftarrow{h}_i^t\right]$, is concatenating forward and backward as the output vector.

Violation encoder.

The input is the sentence describing the violation situation consisting of l words$S_{violation} = \{w_1, w_2, ..., w_l\}$. Four layers learn the representation, the first three layers represent word-level and the last layer represents document-level. With an embedding matrixW_w^V, the sentence is converted to a sequence of vectors$E^V = \{e_1^V, e_2^V, ..., e_k^V\}$, where, $W_w^V \in R^{d\times|V^V|}$, d is the dimension of the word embeddings and $|V^V|$ is the vocabulary size of violation. The second and third layers are fully connected layers which are used to learn the hidden representation of the inspection situation:

$$x_n^v = ReLu(W_2^V \cdot ReLu(W_1^V e_n^V + b_1^V) + b_2^V)$$

where $W_2^V, W_1^V, b_1^V, b_2^V$ are weight matrices and bias vectors in the inspection encoder, x_n^v denoted the n^{th} hidden representation of the word w_n in violation comments.

The fourth layer is a Bidirectional Gate Recurrent Unit (Bi-GRU) network. It takes the hidden word embeddings as input and outputs final cell hidden states as document-level representations. Let $\overrightarrow{h}_n^t$ and $\overleftarrow{h}_n^t$ denote the t^{th} forward and backward hidden layer representation of the wordw_n, which are calculated by:

$$\overrightarrow{h}_n^t = \overrightarrow{GRU}(x_n^v), \overleftarrow{h}_n^t = \overleftarrow{G}RU(x_n^v)$$

the final hidden layer vector of the wordw_k, denoted as$h_{violation} = \left[\overrightarrow{h}_i^t; \overleftarrow{h}_i^t\right]$, is concatenating forward and backward as the output vector.

Violation type encoder.

In contrast with formers, violation type is just a label with very few words, such as "environmental health and safety" and "administrative". Specifically, it is classified after confirming the detailed content of the violation. Here, we design a module containing three layers for the violation type encoder. The inputs of this encoder are the identifiers (I.D. s) of the violation types. The first layer is an embedding layer, which can transform the discrete I.D.s into low-dimensional dense representation vector$e^{vt} \in R^{D_{vt}}$, where D_{vt} is the embedding dimension of the violation type I.D. The second and the third layer are fully connected layers, then the categorical level representation x_n^{vt} of n^{th} violation type is:

$$x_n^{vt} = ReLu(W_2^{vt} \cdot ReLu(W_1^{vt} e_n^{vt} + b_1^{vt}) + b_2^{vt})$$

where $W_2^{vt}, W_1^{vt}, b_1^{vt}, b_2^{vt}$ are weight matrices and bias vectors.

3.2. Violation severity encoder.

The input of this encoder is the categorical label of violation severity, which is also a label with very few words. There are several violation severities, such as the "noted violation", the "outstanding violation", and the "recurrent violation". The inputs are the identifiers (I.D.s) of the violation severity. The first layer is an embedding layer, which can transform the discrete I.D.s into a low-dimensional dense representation vector$e^{vs} \in R^{D_{vs}}$, where D_{vs} is the embedding dimension of the violation severity I.D. The second and the third layer are fully connected layers, then the categorical level representation x_n^{vs} of n^{th} violation severity is:

$$x_n^{vs} = ReLu(W_2^{vs} \cdot ReLu(W_1^{vs} e_n^{vs} + b_1^{vs}) + b_2^{vs})$$

where $W_2^{vs}, W_1^{vs}, b_1^{vs}, b_2^{vs}$ are weight matrices and bias vectors.

3.3. Year encoder

The input of this encoder is originally the numerical label of the year, which is from 2001 to 2019. To process conveniently, we converted it into categorical variables. The inputs are the identifiers (I.D.s) of the year. The first layer is an embedding layer, which can transform the discrete I.D.s into a low-dimensional dense representation vector$e^y \in R^{D_y}$, where D_y is the embedding dimension of the year I.D. The second and the third layer are fully connected layers, then the

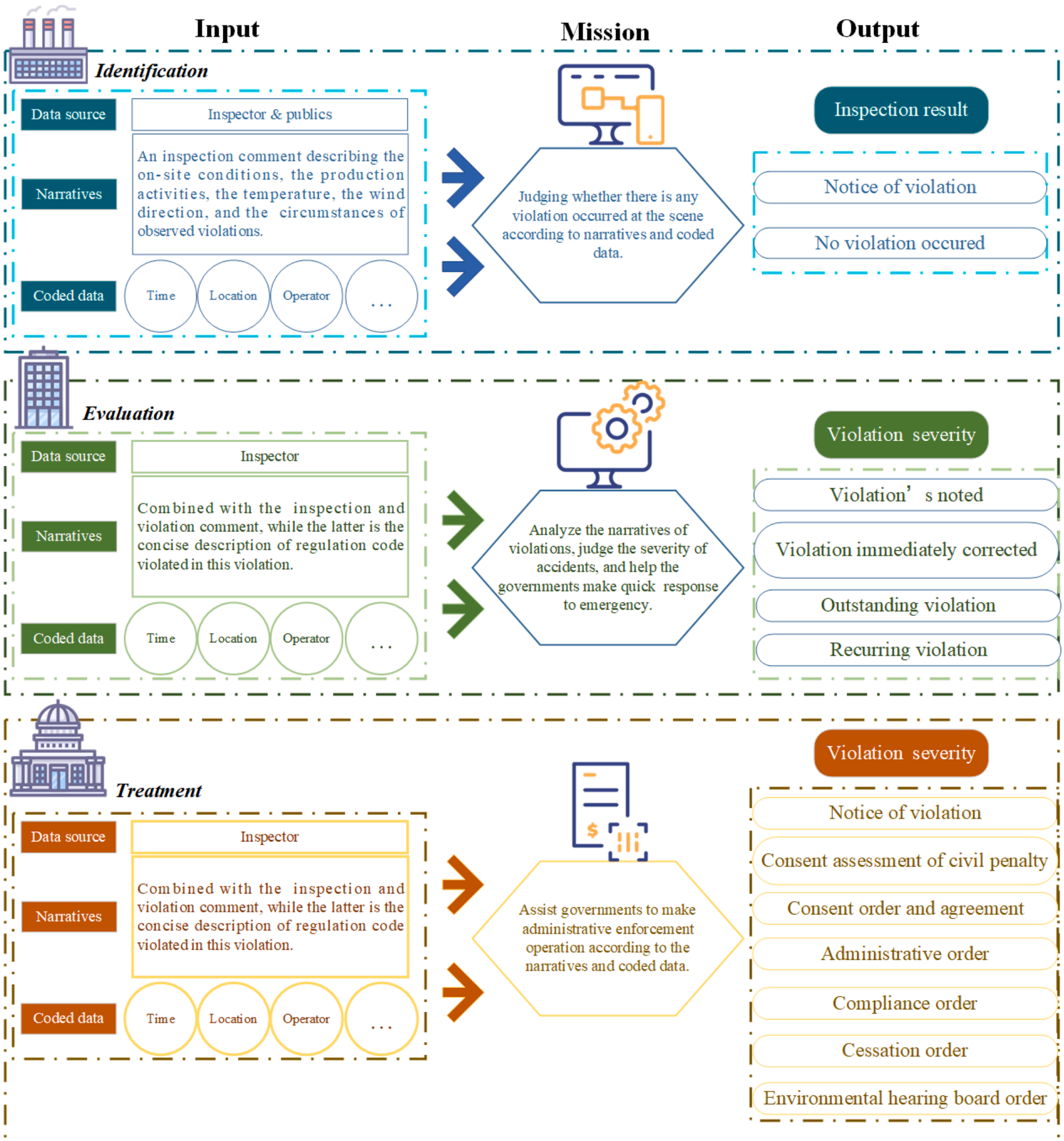

Fig. 3. The framework of this paper.

categorical level representation x_n^y of n^{th} year is:

$$x_n^y = ReLu(W_2^y \cdot ReLu(W_1^y e_n^y + b_1^y) + b_2^y)$$

where $W_2^y, W_1^y, b_1^y, b_2^y$ are weight matrices and bias vectors.

Site encoder.

The inputs are the identifiers (I.D.s) of the site, such as “816699″, ”742198″, and “734718″. The first layer is an embedding layer, which can transform the discrete I.D.s into a low-dimensional dense representation vector$e^s \in R^{D_s}$, where D_s is the embedding dimension of the site I.D. The second and the third layer are fully connected layers, then the categorical level representation of n^{th} site x_n^s is:

$$x_n^s = ReLu(W_2^s \cdot ReLu(W_1^s e_n^s + b_1^s) + b_2^s)$$

where $W_2^s, W_1^s, b_1^s, b_2^s$ are weight matrices and bias vectors.

County encoder.

The input of this encoder is the categorical label of the county, such as “Susquehanna”, “Washington”, and “Potter”. The inputs are the identifiers (I.D.s) of the county. The first layer is an embedding layer, which can transform the discrete I.D.s into a low-dimensional dense representation vector$e^c \in R^{D_c}$, where D_c is the embedding dimension of the county I.D. The second and the third layer are fully connected layers, then the categorical level representation x_n^c of n^{th} county is:

$$x_n^c = ReLu(W_2^c \cdot ReLu(W_1^c e_n^c + b_1^c) + b_2^c)$$

where $W_2^c, W_1^c, b_1^c, b_2^c$ are weight matrices and bias vectors.

Operator encoder.

The input of this encoder is the categorical label of the operator, such as “Chesapeake Appalachia LLC”, “Range Resources Appalachia LLC”, and “Chief Oil & Gas LLC”. The inputs are the identifiers (I.D.s) of the

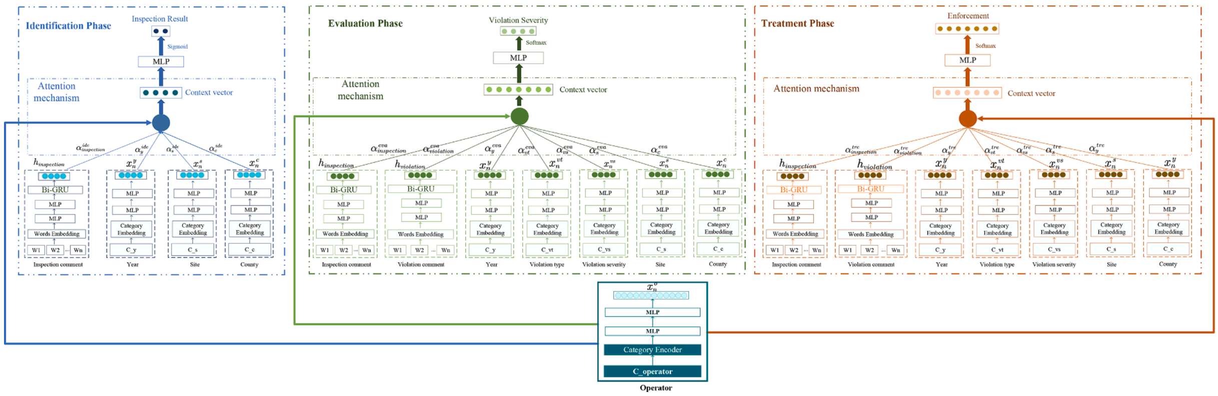

Fig. 4. The attentive neural network framework for violation supervision3.2 Basic components.

operator. The first layer is an embedding layer, which can transform the discrete I.D.s into a low-dimensional dense representation vector$e^o \in R^{D_o}$, where D_o is the embedding dimension of the operator I. D. The second and the third layer are fully connected layers, then the categorical level representation x_n^o of n^{th} operator is:

$$x_n^o = ReLu(W_2^o \cdot ReLu(W_1^o e_n^o + b_1^o) + b_2^o)$$

where $W_2^o, W_1^o, b_1^o, b_2^o$ are weight matrices and bias vectors.

3.4. Identification phase

The first layer is the attention layer, and it produces an attention weight vector and content vector as output. We concatenate the outputs of encoders, including Inspection encoder, Year encoder, Site encoder, and County encoder, into the value tensor in the identification phase (abbreviated as *ide*):$h_V^{ide} = [h_{inspection}; x_n^y; x_n^s; x_n^c]$, and the output of the Operator encoder is the query tensor$h_q^{ide} = x_n^o$. Bahdanau's (Bahdanau et al., 2014) additive style is used to calculate the attention weight of value and query tensor, then this layer generates context vector as output:

$$\alpha_v^{ide} = \frac{\exp(score(h_q^{ide}, h_v^{ide}))}{\sum_v^V \exp(score(h_q^{ide}, h_v^{ide}))}[Attention\_weights]$$

$$c_q^{ide} = \sum_v \alpha_v^{ide} h_v^{ide}[Context\_vector]$$

$$score(h_q^{ide}, h_v^{ide}) = v_a^T \tanh(W_1^{ide} h_q^{ide} + W_2^{ide} h_v^{ide})[Attention\_score]$$

where v and q denotes the subscript of value and query tensor respectively, V denotes the set of all subscripts of value tensors, in this phase,$V = [inspection, year, site, county]$.

After that, two fully connected layers are used to produce the probabilityp^{ide}, which if the violation occurs, in the identification phase:

$$p^{ide} = sigmoid(W_4^{ide} ReLu(W_3^{ide} c_q^{ide} + b_3^{ide}) + b_4^{ide})$$

where $W_4^{ide}, W_3^{ide}, b_3^{ide}, b_4^{ide}$ are weight matrices and bias vectors.

3.5. Evaluation phase

The first layer is the attention layer, and we concatenate the outputs of encoders, including Inspection encoder, Violation encoder, Violation type encoder, Violation severity encoder, Year encoder, Site encoder, and County encoder as value tensor in the evaluation phase (abbreviated as *eva*):$h_V^{eva} = [h_{inspection}; h_{violation}; x_n^{vt}; x_n^{vs}; x_n^y; x_n^s; x_n^c]$, and the output of Operator encoder is the query tensor$h_q^{eva} = x_n^o$. α_v^{eva} denotes the attention weight of value and query tensor, then this layer generates context vector c_q^{eva} as output:

$$\alpha_v^{eva} = \frac{\exp(score(h_q^{eva}, h_v^{eva}))}{\sum_v^V \exp(score(h_q^{eva}, h_v^{eva}))}[Attention\_weights]$$

$$c_q^{eva} = \sum_v \alpha_v^{eva} h_v^{eva}[Context\_vector]$$

$$score(h_q^{eva}, h_v^{eva}) = v_a^T \tanh(W_1^{eva} h_q^{eva} + W_2^{eva} h_v^{eva})[Attention\_score]$$

where v and q denotes the subscript of value and query tensor respectively, V denotes the set of all subscripts of value tensors, in this phase,$V = [inspection, violation, vt, vs, year, site, county]$.

Two fully connected layers are used to produce the probability p^{eva} of violation severities:

$$p^{eva} = softmax(W_4^{eva} ReLu(W_3^{eva} c_q^{eva} + b_3^{eva}) + b_4^{eva})$$

where $W_4^{eva}, W_3^{eva}, b_3^{eva}, b_4^{eva}$ are weight matrices and bias vectors.

3.6. Treatment phase

The first layer is the attention layer. We concatenate the outputs of encoders, including Inspection encoder, Violation encoder, Violation type encoder, Violation severity encoder, Year encoder, Site encoder, and County encoder, into value tensor in the treatment phase (abbreviated as *tre*):$h_V^{tre} = [h_{inspection}; h_{violation}; x_n^{vt}; x_n^{vs}; x_n^y; x_n^s; x_n^c]$, while the output of Operator encoder is the query tensor:$h_q^{tre} = x_n^o$. α_v^{tre} denotes the attention weight of value and query tensor, then this layer generates context vector c_q^{tre} as output:

$$\alpha_v^{tre} = \frac{\exp(score(h_q^{tre}, h_v^{tre}))}{\sum_v^V \exp(score(h_q^{tre}, h_v^{tre}))}[Attention\_weights]$$

$$c_q^{tre} = \sum_v \alpha_v^{tre} h_v^{tre}[Context\_vector]$$

$$score(h_q^{tre}, h_v^{tre}) = v_a^T \tanh(W_1^{tre} h_q^{tre} + W_2^{tre} h_v^{tre})[Attention\_score]$$

where v and q denotes the single subscript of value and query tensor respectively, V denotes the set of all subscripts of value tensors, in this phase,$V = [inspection, violation, vt, vs, year, site, county]$.

Two fully connected layers are used to produce the probability p^{tre} of enforcement classes in our treatment phase:

$$p^{tre} = softmax(W_4^{tre} ReLu(W_3^{tre} c_q^{tre} + b_3^{tre}) + b_4^{tre})$$

where $W_4^{tre}, W_3^{tre}, b_3^{tre}, b_4^{tre}$ are weight matrices and bias vectors.

3.7. Model training and test stage

During the training process, the loss function is defined as the cross-entropy of the prediction probability and the ground truth for each phase. It can be calculated by:

$$l(\hat{y}) = -\sum_{x_j \in K}^{x} y_j \log(\hat{y}_j)$$

The test set, including binary and multilabel, $\hat{y} = \left\{\hat{y}_1, \hat{y}_2, \dots, \hat{y}_{|K|}\right\}$ is the prediction probabilities over each phase in$K$. Where y_j is the ground truth probability distribution of test setx_j. Specifically, if x_j is the positive label, then$y_j = 1$, otherwise$y_j = 0$.

By minimizing the loss function $l(\hat{y})$ via Adam optimizer, all parameters can be tuned in the training process.

4. Experiments

4.1. Experiment setup

4.1.1. Data preprocessing

Data preprocessing is the most critical work before constructing neural networks. Some compliance reports are duplicated, and the narrative texts need to be transformed into structure embeddings for the following work. There are two essential steps in data preprocessing, that is, narrative data preprocessing and imbalanced data preprocessing.

Source.

We obtained a total of 145,123 the UOG violation samples from the Pennsylvania Department of Environmental Protection (website: https://www.dep.pa.gov/Pages/default.aspx), which assigned the oil type as “unconventional” and time from 2001 to 2019.

Data description.

There are three phases in our proposed neural attentive networks, the inputs and outputs are varied from each stage, which are, Identification, Evaluation, and Treatment. (1) In the Identification phase, it is formulated as multi-inputs and single output questions. Our proposed model needs to identify whether there exists a violation mainly according to inspection narrative reviews; (2) while in the Evaluation phase, it is formulated to evaluate the severity of violations according to narrative texts and on-site coded data; (3) In the Treatment phase, it needs to decide which enforcement should be taken according to violation severity, situation, and location. The detailed parameters and description are listed in Table 1.

Narratives preprocessing.

Narratives data preprocessing in our study included the following three essential steps, that is the tokenization, lowercase, and stop-words removal, as shown in Fig. 5:

- Tokenization: tokenization is converting complex text (namely the corpus) into a set of words called tokens.
- Lowercase: Convert all uppercase characters in the text to lowercase characters.
- Stop-words removal: this step eliminates stop = words that provide no valuable information for violation analysis, such as pronouns, prepositions, symbols (@,!,$,#, etc.). Generally speaking, it can be roughly divided into two categories: the first is the functional words, such as “the”, “is”, “which”, and “on”, etc. While another type is verbs, such as “want”, “do”, “make”, etc., these kinds of words have little practical meaning, we will remove these kinds of stop-words in this preprocessing stage.

Imbalanced data.

Unbalanced data is the most significant feature in our data set. In the Identification stage, we want our model to obtain the ability to judge whether a violation occurred (especially the attributes and attention weights when there is a violation). However, after deleting duplicated reports, we noted that the dependent variable ’violation’ representation had a significantly small number of ’no violation’ representations as 5 % vs 95 %. Unbalanced data results in the model being lopsided towards the majority class (’no violation’), the attributes and attention weights learned by neural networks can also be distorted during the training stage. Since there is no best way to handle the unbalanced dataset in all cases, after experimenting with standard methods, that is, oversampling, under-sampling, and a mixture of over and under-sampling, we decided to apply over-sampling in identification and treatment stages for the following reasons: (1) Under-sampling will almost eliminate the 95 % samples in the biggest label set, those missing representations of

Table 1
The data description in three phases.

Phase	Independent Variables	Type	Description	Dependent variables
Identification	Inspection comment	Textual	A narrative text described the inspection situation at the site.	Result of Inspection.
	Year	Numerical	The year of inspection.	
	Site	Categorical	The name of the site.	
	County	Categorical	The name of the county.	
	Operator	Categorical	The name of the site operator.	
Evaluation	Inspection comment	Textual	A narrative text described the inspection situation at the site.	The severity of the violation.
	Violation comment	Textual	A narrative text described the violation situation at the site.	
	Year	Numerical	The year of violation.	
	Site	Categorical	The name of the site.	
	County	Categorical	The name of the county.	
	Operator	Categorical	The name of the site operator.	
Treatment	Inspection comment	Textual	A narrative text described the inspection situation at the site.	Enforcement.
	Violation comment	Textual	A narrative text described the violation situation at the site.	
	Year	Numerical	The year of violation.	
	Violation type	Categorical	The types of violation that is, environmental or administrative.	
	Violation severity	Textual	A narrative text described the severity of the violation.	
	Site	Categorical	The name of the site.	
	County	Categorical	The name of the county.	
	Operator	Categorical	The name of the site operator.	

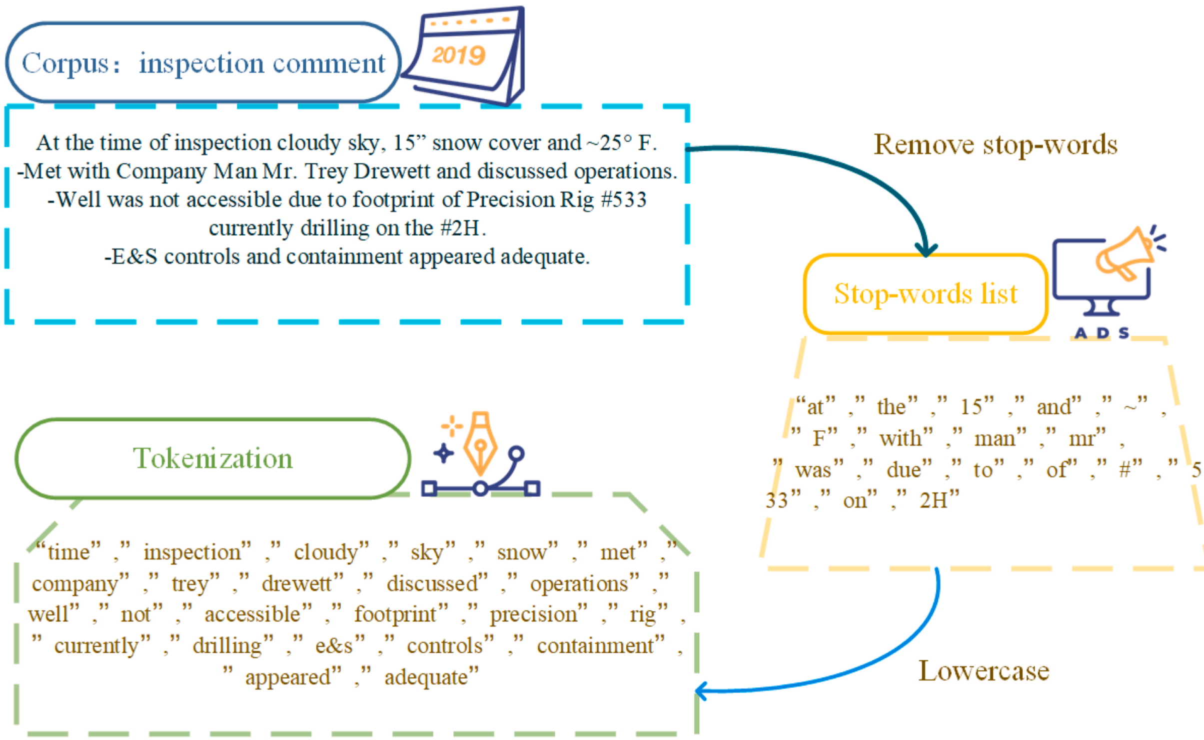

Fig. 5. Narratives preprocessing.

narrative text are fatal for the model to learn the attention weights among different indicators in reality world; (2) The traditional mixture of over and under sampling cannot fit in narrative text data, it is challenging to generate a 'noised' narrative text without distortion of semantic meanings; (3) While we will try to maintain the original ratio of labels and make an appreciable increase in small-scale samples. Statistics of these three datasets are given in Table 2.Table 3..

4.1.2. Hyperparameters

We will illustrate the hyperparameter settings in our neural networks in this section. Dims of the word and category embedding are set as 128. In the identification model, the size of Bi-LSTM equals 128, and the size of the fully connected layer also equals 256. For Logistic Regression, we utilized l2 regulation in the loss function. The inverse of regularization strength equals 0.00001. For SVM, we utilized the Gaussian kernel as a kernel function.

4.1.3. Baseline methods

We compared the proposed framework with three classic machine learning methods (i.e., SVM, L.R., Random Forests) and three state-of-the-art neural classifiers (i.e., GRU, Bi-GRU, LSTM). These algorithms are well established and carefully described in machine learning textbooks (e.g., Zhihua Zhou,2021). Thus, the remainder of this section only provides brief descriptions.

- SVM: Support Vector Machine (SVM) is a generalized linear classifier that performs binary data classification according to supervised learning. Its decision boundary is the maximum-margin hyperplane for solving the learning sample.

Table 2
Statistics of three datasets.

Dataset	Train (60 %)	Dev (10 %)	Test (30 %)
Identification	32,051	3561	15,263
Evaluation	6578	730	3132
Treatment	5166	574	2461

Table 3
Performance comparation of different methods.

Phase	Model	Precision	Recall	F1-score
Identification	GRU + Attention	0.9809	0.9806	0.9807
	LSTM + Attention	0.9604	0.9593	0.9598
	Bi-GRU + Attention*	**0.9822**	**0.9814**	**0.9818**
	Bi-LSTM + Attention	0.9760	0.9756	0.9758
	SVM	0.8631	0.8629	0.8630
	Random Forests	0.6405	0.8003	0.7115
	Logistic Regression	0.7393	0.7983	0.7676
Evaluation	GRU + Attention	0.8778	0.8799	0.8788
	LSTM + Attention	0.8859	0.8818	0.8838
	Bi-GRU + Attention	**0.9147**	**0.9153**	**0.9150**
	Bi-LSTM + Attention	0.8850	0.8866	0.8857
	Random Forest	0.7037	0.7314	0.7172
	SVM	0.8761	0.8588	0.8673
	Logistic Regression	0.7669	0.7905	0.7785
Treatment	GRU + Attention	0.6942	0.6972	0.6956
	LSTM + Attention	0.7053	0.6948	0.7001
	Bi-GRU + Attention*	**0.7019**	**0.7011**	**0.7015**
	Bi-LSTM + Attention	0.7173	0.6827	0.6995
	SVM	**0.6705**	**0.7556**	**0.7105**
	Random Forests	0.5334	0.7303	0.6165
	Logistic Regression	0.6492	0.7351	0.6894

- L.R.: Logistic regression algorithm predicts the probability of occurrence of an event by fitting data to a logit function (Goh & Ubeynarayana, 2017).
- Random Forests: a random forest classifier constitutes a set of decision trees. Each tree in the forest predicts its final class label. The collection of trees then votes for the most popular class as the final class label (Goh & Ubeynarayana, 2017).

4.2. Performance metrics

We applied three widely used metrics to evaluate the efficiency of our proposed neural networks. That is, weighted-Precision, weighted-Recall, and F1-score.

$$Precision_{weighted} = \sum_{i=1}^{n} \frac{TP_i}{TP_i + FP_i} \times \omega_i$$

$$Recall_{weighted} = \sum_{i=1}^{n} \frac{TP_i}{TP_i + FN_i} \times \omega_i$$

$$F1 = \frac{2 \times precision_{weighted} \times recall_{weighted}}{precision_{weighted} + recall_{weighted}}$$

where TP_i denotes the true positives of category i, FP_i denotes the false positives of category$i$, FN_i denotes the false negatives of category$i$, ω_i denotes the weight of category$i$, there are $i = (1, ..., n)$ categories. From the perspective of prediction results, precision describes how many of the predicted positive examples are true positives; from the perspective of real results, recall describes how many true positives in the test set are correctly predicted. While they are usually a pair of contradictory metrics. The higher the precision, the lower the recall. And F1 characterizes the comprehensive performance in terms of precision and recall, it is the harmonic average of the two.

4.3. Training

This section describes the training regime for our work.

Training batch.

We set 60 % as the train set, 10 % as the validation set, and 30 % as the test size, the batch size is set to 256.

Training environment.

The proposed attentive neural networks and all the compared neural-based models are defined and trained on Google Collaboratory and implemented in TensorFlow 2.

Optimizer.

Adam (Kingma & Ba, 2014) are applied in optimizing the model, and the learning rate equals 0.001.

Initialization.

For the fully connected layers, the activation is the 'ReLU' function:$f(x) = \max(0, w^T x + b)$, the kernel initializer is the glorot uniform, and the bias initializer is zeros.

Regularization.

We applied early stopping in deep neural networks during training to avoid overfitting. The early stopping patience is set to 10, and the epochs are set to 50.

5. Results

The effectiveness of our proposed attentive neural networks is discussed in three ways: (1) First, the performance of our proposed framework is discussed with state-of-art neural networks and widely used machine learning methods; (2) we further analyzed the effectiveness of the attention mechanism inside our neural networks in different phases; (3) finally, we visualized attention weight inside our model to interpret the importance of factors.

5.1. Our proposed work performance

In this section, we compare the overall performance between our proposed framework and benchmarks. The aim is to select the most time-saving and optimal model in three phases.

The results are shown in Table 2. There are four deep neural networks: GRU + Attention, Bi-GRU + Attention, LSTM + Attention, and Bi-LSTM + Attention; with three widely used machine learning methods: Support vector machine (SVM), Random Forests, and Logistic Regression. The achieved results show that our proposed framework performed best in the way of Bi-GRU + Attention. Specifically, we have several observations.

First, BI-GRU performed the best among other benchmarks in all stages. Its F1-scores are 98.18 %, 94.5 %, and 70.15 % respectively. It is not surprising that Bi-GRU performed better than GRU, since Bi-GRU considers bidirectional information in narrative texts, while not significantly improving performance. Unexpectedly, the performance of Bi-LSTM in the three stages is not satisfactory. For example, in the identification phase, its performance is significantly lower than the other three deep neural networks and SVM. We argue that Bi-LSTM overfits the text information on the training set, which leads to poor performance on the test set.

Secondly, SVM performed best among other classic machine learning algorithms, followed by LR. and RF. Finally, it is worth noting that the performances of some deep neural networks are not as accurate as SVM. For example, in the treatment stage, the F1-score of SVM is even better than that of Bi-GRU + Attention. These results indicated that we could continue to explore the application of machine learning algorithms in text mining and classification programs.

5.2. Effectiveness of attention network

In this section, we further analyze the effectiveness of attention mechanisms in deep neural networks, Figs. 6-8 shows the performance of neural networks with/without attention mechanisms in three stages. The followings are observations and discussions.

First, we must admit that the neural networks with the attention mechanism do not perform well in all phases. But it is worth noting that in the identification stage, the attention mechanism dramatically improves the performance of all neural networks. During the experiment, we found that the number of small-scale samples most affected the performance of multi-classification models. However, there are some defects in our original small-scale data. For example, the duplicate narrative records are shared with different labels due to the careless negligence of staff. The attention mechanism amplified these defects, which makes them perform poorly.

Moreover, among all neural networks without attention mechanism in three stages, the Bi-LSTM achieved the best performance. In the evaluation and treatment stages, the performance of Bi-LSTM surpasses the Bi-GRU + Attention, with its F1-score as 0.9215 and 0.7052, respectively. Additionally, as recall and precision are a pair of contradictive metrics, the recall is higher than the precision of all models in three stages, which indicates that our model can better capture the characteristics of true positives.

5.3. Visualization of attention weights

In this section, we further analyze factors' importance based on visualizing attention weights inside neural networks. We applied box plots to represent the data features, while the yellow horizontal line represents the median, and the green dot represents the mean.

5.3.1. Visualize the attention weights in the identification phase

This section visualizes three different attention weights according to the inspection results in Table 4.

First, Fig. 9**(a)** describes the attention weights for the whole samples, including "***no violation***" and "***notice of violation***", two consequences. The "inspection comment" gains the highest attention weight, followed by the "site", "year", and "county". Specifically, the median of the "inspection comment" is 0.28, which confirms that narratives are the most significant indicator in the identification phase. Among other indicators, the average attention weights of the "site" are slightly higher than those of the "year" and "county", which indicated that different sites have more implicated geographical features than the county. In contrast, the geographical feature (site, county) is more significant compared with time features (year).

Second, Fig. 9**(b) and (c)** display the attention weights of only "***notice of violation***" and "***no violation***", respectively. It's noted that the "inspection comment" still gains the highest attention weight in two

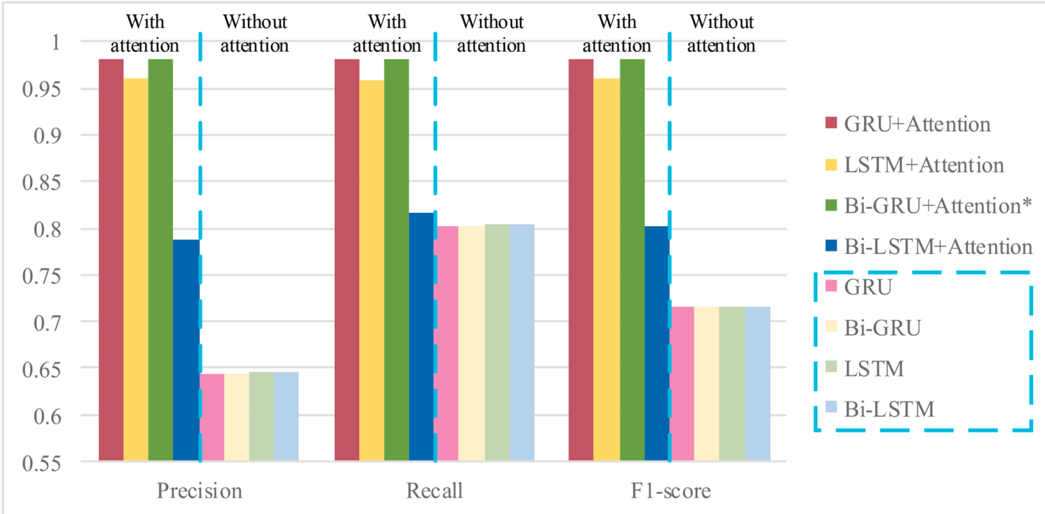

Fig. 6. Identification performance with/without attention.

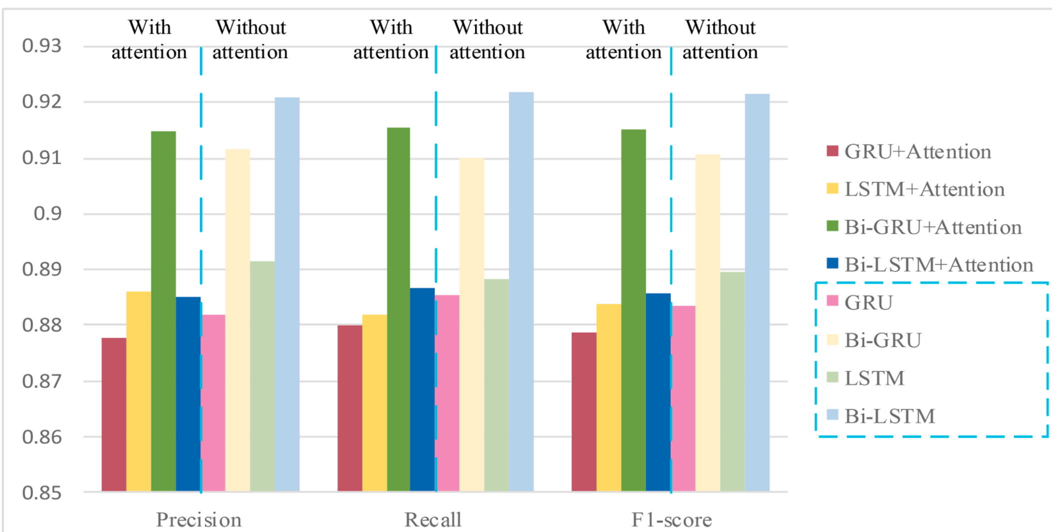

Fig. 7. Evaluation performance with/without attention.

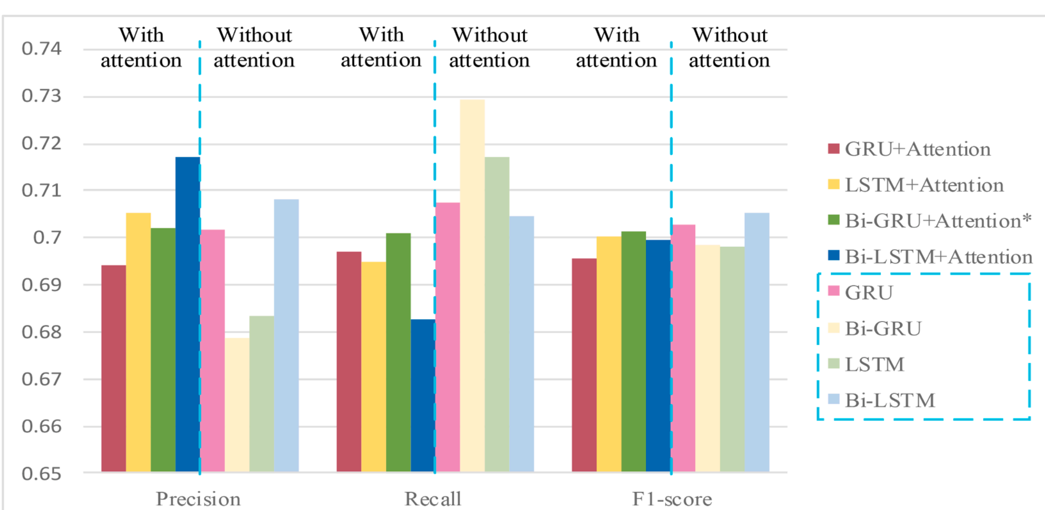

Fig. 8. Treatment performance with/without attention.

Table 4
Samples on labels of the independent variable.

Independent variable	Labels	Samples
Inspection result	No violation	40,876
	Notice of violation	10,000

cases, once again emphasizing that our model can extract powerful document-level representations from narrative texts. While there are some differences in the attention weights of the "year", "site", and "county" between these two samples, they are slightly lower in the case of "***notice of violation***". The "site" gains the highest attention weight, followed by the "year" and "county". Specifically, the medians are around 0.2332, 0.2264, and 0.2226.

5.3.2. Visualize the attention weights in the evaluation phase

We visualize the attention weights in the evaluation phase according to the types and severity of violations summarized in Table 5.

The attention weights of labels are shown in Fig. 10, and the following are observations and discussions. First, we noticed that the "site" gains the highest attention weight (mean and medians), followed by the "inspection comments" (Fig. 10**(a)**). It implicates that geographic feature is one of the essential factors affecting the severity of violations.

Secondly, we noticed that the attention weight of the "violation comment" is lower than the "inspection comment" in all cases. In other words, the content recorded in the "inspection comment" is more helpful in defining the severity of violations, the reason is that the "inspection comment" is more detailed than "violation comment" (Fig. 1). The corpus of the "violation comments" has a total of 1288 words, while the "inspection comments" contain 9779 words, which is five times that of the former. Significantly, the "inspection comment" describes the weather, wind direction, temperature, and on-site situation, while the "violation comment" describes the regulation concisely. Hence, our model can learn more beneficial document-level representations from inspection comments compared to violation comments.

Finally, we are concerned about two serious violations: the "***outstanding violations***" and "***recurring violations***" (Fig. 10**(c) and 10 (e)**). An outstanding violation is an unresolved citation reported to the failure to solve, while a recurring violation refers to two or more similar incidents at this site. They both show a commonality: the "site" always gains the highest attention weight. It is also consistent with the previous conjecture that certain sites cause significant differences in the degree of environmental and human health damage due to their specific geographic features.

5.3.3. Visualize the attention weights in the Treatment phase

We visualize the attention weights in the Treatment phase according to labels of enforcement (Table 6).

The attention weights are shown in Fig. 11. Following are some observations and discussions.

First, the "site" gains the highest attention weight from the perspective of all samples, the "***notice of violation***", and the "***consent assessment of civil penalty***" samples (Fig. 11**(a), (c), (d)**). It implies that the "site" is an essential factor in the following respects: different sites contain latent threats based on geological features. Government officers have different attitudes towards environmental violations. The complex relationship between geological features and environmental regulations is one of the main reasons that make the site the most important feature. Therefore, it can significantly help our model to extract representation to determine the enforcement. Unexpectedly, inspection and violation comments obtain low attention weights in these cases. Through the review of the original data and error experiments, we reckon that there is one fundamental reason behind this phenomenon: it occasionally happened in our original data that several enforcement labels share the joint inspection and violation comment, which makes our model cannot extract powerful narrative features to classify enforcements effectively.

It cannot be ignored that the "year" gains a high attention weight (median), especially in the "***consent order***" and "***cessation order***" samples (Fig. 11**(b), (e)**). We reckon that there are two reasons: on the one hand, these samples pertain to environmental violations with obvious damage, which are usually small-scale and highly connected with time features; on the other hand, the stringency of environmental

Table 5
Samples on labels of the independent variable.

Independent variable	Labels	Samples
Violation severity	Violation's noted	7489
	Outstanding violation	1878
	Violation's noted and immediately corrected	985
	Recurring violation	88

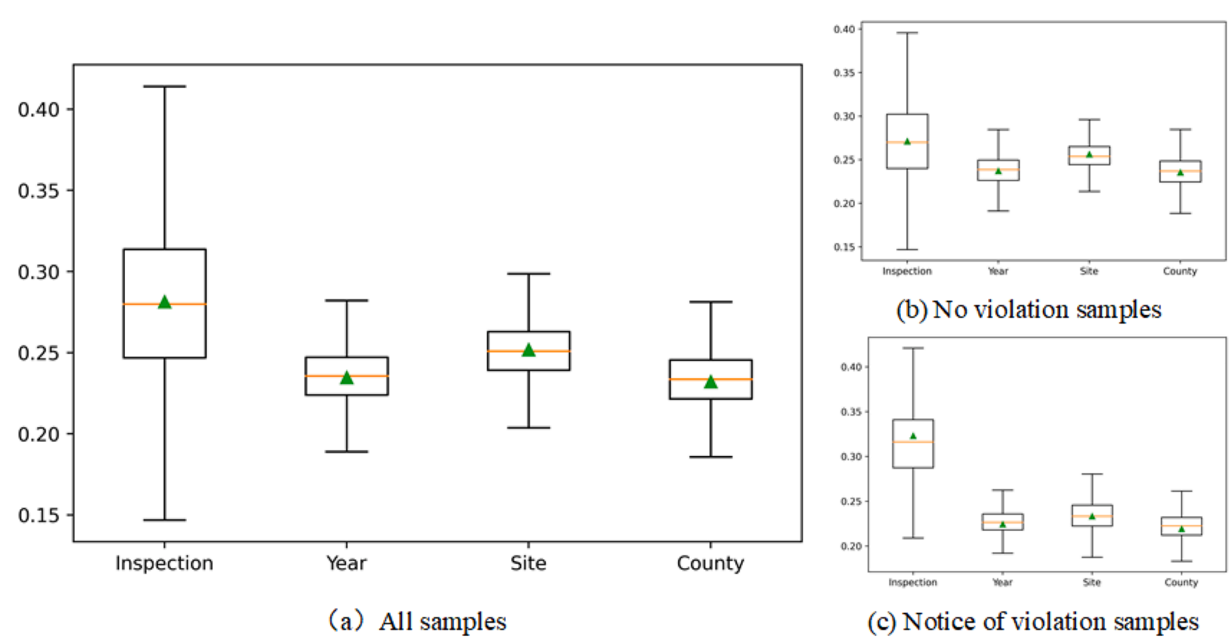

Fig. 9. The attention weights in the Identification phase.

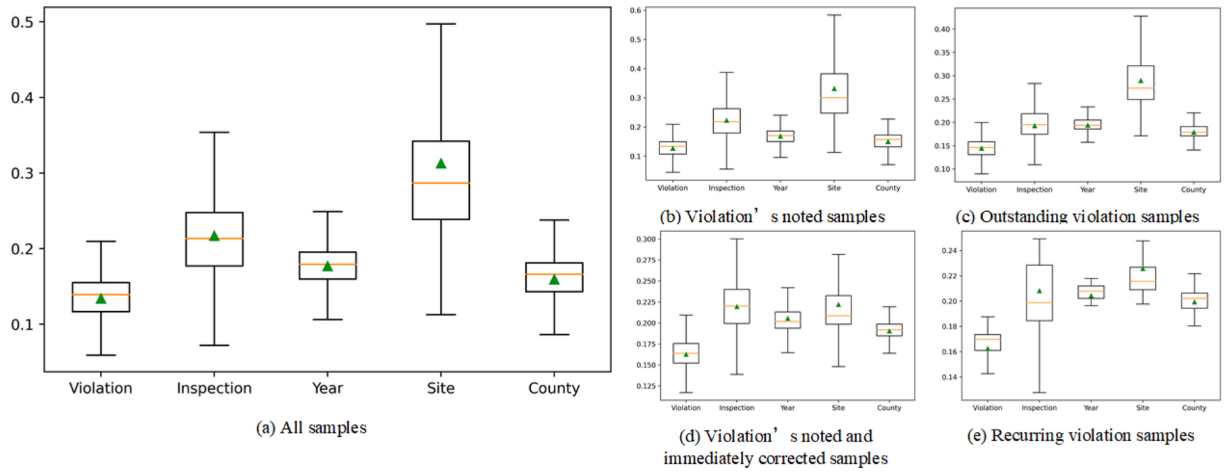

Fig. 10. The attention weight in the evaluation phase.

Table 6
Samples on labels of the independent variable.

Independent variable	Labels	Samples
Enforcement	Notice of violation	5001
	Consent assessment of civil penalty	1280
	Consent order and agreement	226
	Administrative order	180
	Compliance order	100
	Cessation order	50
	Environmental hearing board order	20

regulations in different years has led to various enforcements.

5.4. Computation time

We present the training times (in average per epoch) and prediction time (evaluate time in the test set) of our proposed framework and other state-of-art deep neural networks (with/without attention) on the UOG compliance reports dataset. The results are shown in Table 7. GRU is the most efficient but with poor classification performance, it can't form the powerful document-level representations for follow-up tasks. Although, Bi-GRU is less efficient as it includes the bidirectional GRU unit. Interestingly, the neural network with the attention mechanism has a shorter training time compared to those without it. We reckon that the attention mechanism is able to generate a beneficial context vector and play a role in regularization to a certain extent, making convergence faster.

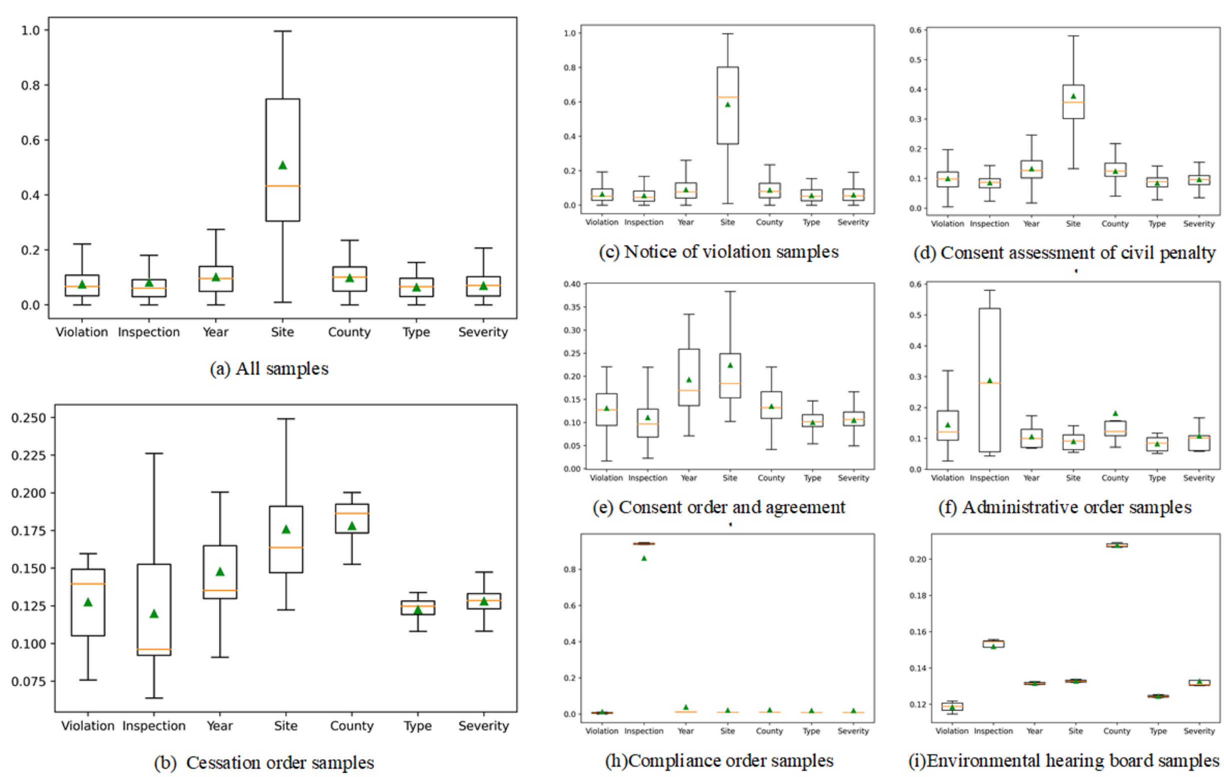

Fig. 11. The attention weights in the Treatment phase.

Table 7
The computation time.

Phase	Model	Training (s/epoch)	Inference (s)
Identification	GRU + Attention	95	22
	LSTM + Attention	119	23
	Bi-GRU + Attention*	179	31
	Bi-LSTM + Attention	208	42
	GRU	110	20
	LSTM	111	21
	Bi-GRU	206	29
	Bi-LSTM	215	36
Evaluation	GRU + Attention	33	6
	LSTM + Attention	38	23
	Bi-GRU + Attention*	60	19
	Bi-LSTM + Attention	80	26
	GRU	35	12
	LSTM	41	11
	Bi-GRU	68	13
	Bi-LSTM	78	21
Treatment	GRU + Attention	23	7
	LSTM + Attention	29	5
	Bi-GRU + Attention*	45	7
	Bi-LSTM + Attention	52	9
	GRU	23	3
	LSTM	26	6
	Bi-GRU	45	8
	Bi-LSTM	53	10

6. Conclusions and discussions

We proposed and tested an attention-based neural network to support decision-making with respect to UOG on-site emergency responses. We applied Bi-GRU combined with the attention mechanism to extract document-level representations and generate context vectors for follow-up tasks, especially aimed at unstructured data. Specifically, we first split the UOG life-cycle violations into three phases, that is, the identification, evaluation, and treatment phase, to support decision making and obtain practical implications based on a real-world dataset. To the best of our knowledge, it is the first work that introduced the attention mechanism into UOG on-site emergency responses system.

6.1. Practical implications

There are several observations obtained from visualizing the attention weights in different phases. After thorough consideration, we concluded two critical practical implications both for governments and practitioners:

First, geographical characteristics (site) are a powerful indicator for supporting decision-making in the identification, evaluation, and treatment phases. Especially in the identification phase, the site gains the highest attention among the “no violation” cases, which was significantly higher than in the “notice of violation” cases. It indicates that the geological location is an inherent factor that predominantly affects the latent environmental damage. In other words, there is an urgent need for governments to implement different inspection strategies according to sites’ geological features, which also benefits governments to save human labor and financial expenditures.

Second, we find that year is a significant feature that affects the type of violations in the evaluation phase. It inspired an interesting conjecture: the type of violation is highly related to the climate in that year. For example, peculiar climatic conditions like continuous precipitation or extreme heat would lead to on-site violations. Assuming that the geological, temporal, and climatic chrematistics can be obtained, it will significantly improve the prediction accuracy of types of on-site violations.

6.2. Limitations and future works

The essential limitations in this paper can be concluded in the following aspects:

- The same and duplicated narrative inspection/violation comments shared different labels in UOG compliance reports, which greatly affects the performance of multi-label classifications in our task. It was due to several rule-breaking activities that co-occurred in a violation accident, while compliance reports only record the overall inspection comment once. Additionally, the enforcement status of the same violation is changed. At the same time, the original record is still saved in compliance reports, which leads to the same violation comment shared with different enforcements.
- The lack of guidelines for recording information in compliance reports is fatal for understanding the complexity of the violation event and conducting in-depth narrative analysis (McKenzie et al., 2010). For example, the lack of details in inspection/violation comments led to errors in interpretations and made our models hard to extract useful information in some cases. Such defects are even more severe in small-scale violations, which leads to underestimating accidents on a small scale with irreparable harm to the environment and human health.

We reckon that interpretable neural networks still have a broad exploration space in applying UOG accident analysis and prevention systems for future work. We intend to focus on combining word-level expression with attention mechanism, some state-of-art embedding methods such as word2vector (Ali et al., 2021), then combined it with deep document-level representations to explore more interpretable mechanisms.

Ethics approval and consent to participate.

Not applicable.

Consent for publication.

Not applicable.

Availability of data and materials.

The datasets analyzed during the current study are available in the Department of Environmental protection, the official Pennsylvania government website repository, [https://www.dep.pa.gov].

Funding.

This research work was supported by the National Natural Science Foundation of China under Grant Nos. 71774130.

CRediT authorship contribution statement

Dan Bi: Conceptualization, Methodology, Software, Validation, Formal analysis, Writing – original draft. **Ju-e Guo:** Resources, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

Abdat, F., Leclercq, S., Cuny, X., & Tissot, C. (2014). Sep). Extracting recurrent scenarios from narrative texts using a Bayesian network: Application to serious occupational accidents with movement disturbance. *Accident Analysis and Prevention, 70*, 155–166. https://doi.org/10.1016/j.aap.2014.04.004

Abualfaraj, N., Olson, M. S., Gurian, P. L., De Roos, A., & Gross-Davis, C. A. (2016). Oct). Statistical analysis of compliance violations for natural gas wells in Pennsylvania. *Energy Policy, 97*, 421–428. https://doi.org/10.1016/j.enpol.2016.07.051

Alammar, J (2018). The Illustrated Transformer [Blog post]. Retrieved from https://jalammar.github.io/illustrated-transformer/.

Ali, F., Ali, A., Imran, M., Naqvi, R. A., Siddiqi, M. H., & Kwak, K.-S. (2021, 2021/03/01/). Traffic accident detection and condition analysis based on social networking data. Accident Analysis & Prevention, 151, 105973.https://doi.org/https://doi.org/10.1016/j.aap.2021.105973.

Bahdanau, D., Cho, K., & Bengio, Y. (2014, 2014/9//). Neural Machine Translation by Jointly Learning to Align and Translate. http://arxiv.org/abs/1409.0473.

Bi, D., Guo, J.-e., Zhao, E., Sun, S., & Wang, S. (2021, 2021/11/19). Identifying environmental and health threats in unconventional oil and gas violations: evidence from Pennsylvania compliance reports. Environmental Science and Pollution Research. https://doi.org/10.1007/s11356-021-17500-8.

Deziel, N. C., Brokovich, E., Grotto, I., Clark, C. J., Barnett-Itzhaki, Z., Broday, D., & Agay-Shay, K. (2020, 2020/03/01/). Unconventional oil and gas development and health outcomes: A scoping review of the epidemiological research. Environmental Research, 182, 109124. https://doi.org/https://doi.org/10.1016/j.envres.2020.109124.

Esfahlani, S. S., Shirvani, H., Butt, J., Mirzaee, I., & Esfahlani, K. S. (2022). 2022/9//). Machine Learning role in clinical decision-making: Neuro-rehabilitation video game. *Expert Systems with Applications, 201*. https://doi.org/10.1016/j.eswa.2022.117165

Goh, Y. M., & Ubeynarayana, C. U. (2017). 2017/11/01/). Construction accident narrative classification: An evaluation of text mining techniques. *Accident Analysis & Prevention, 108*, 122–130. https://doi.org/10.1016/j.aap.2017.08.026.

Guo, M., Xu, Y., & Chen, Y. D. (2019). Environmental enforcement and compliance in Pennsylvania's Marcellus shale gas development. *Resources Conservation and Recycling, 144*, 24–31. https://doi.org/10.1016/j.resconrec.2019.01.006

Jacobé de Naurois, C., Bourdin, C., Stratulat, A., Diaz, E., & Vercher, J.-L. (2019). 2019/05/01/). Detection and prediction of driver drowsiness using artificial neural network models. *Accident Analysis & Prevention, 126*, 95–104. https://doi.org/10.1016/j.aap.2017.11.038.

Jia, N., Chen, J., & Wang, R. (2022). 2022/9//). An attention-based convolutional neural network for recipe recommendation. *Expert Systems with Applications, 201*. https://doi.org/10.1016/j.eswa.2022.116979

Kim, B.-C., & Oliver, M. E. (2017). Nov). Taming drillers through legislative action: Evidence from Pennsylvania's shale gas industry. *Resource and Energy Economics, 50*, 15–35. https://doi.org/10.1016/j.reseneeco.2017.06.002

Kingma, D. P., & Ba, J. (2014, 2014/12//). Adam: A Method for Stochastic Optimization. http://arxiv.org/abs/1412.6980.

Kuwayama, Y., Roeshot, S., Krupnick, A., Richardson, N., & Mares, J. (2017). Feb). Risks and mitigation options for on-site storage of wastewater from shale gas and tight oil development. *Energy Policy, 101*, 582–593. https://doi.org/10.1016/j.enpol.2016.11.016

Kwayu, K. M., Kwigizile, V., Lee, K., & Oh, J.-S. (2021, 2021/02/01/). Discovering latent themes in traffic fatal crash narratives using text mining analytics and network topology. Accident Analysis & Prevention, 150, 105899. https://doi.org/https://doi.org/10.1016/j.aap.2020.105899.

Ma, J., Chan, J., Rajasegarar, S., & Leckie, C. (2022, 2022/4//). Multi-attention graph neural networks for city-wide bus travel time estimation using limited data. Expert Systems with Applications, 117057-117057. https://doi.org/10.1016/j.eswa.2022.117057.

Manda, A. K., Heath, J. L., Klein, W. A., Griffin, M. T., & Montz, B. E. (2014). Sep 1). Evolution of multi-well pad development and influence of well pads on environmental violations and wastewater volumes in the Marcellus shale (USA). *Journal of Environmental Management, 142*, 36–45. https://doi.org/10.1016/j.jenvman.2014.04.011

McKenzie, K., Scott, D. A., Campbell, M. A., & McClure, R. J. (2010, 2010/03/01/). The use of narrative text for injury surveillance research: A systematic review. Accident Analysis & Prevention, 42(2), 354-363. https://doi.org/https://doi.org/10.1016/j.aap.2009.09.020.

Occhipinti, A., Rogers, L., & Angione, C. (2022). 2022/9//). A pipeline and comparative study of 12 machine learning models for text classification. *Expert Systems with Applications, 201*. https://doi.org/10.1016/j.eswa.2022.117193

Olmstead, S. M., Muehlenbachs, L. A., Shih, J.-S., Chu, Z., & Krupnick, A. J. (2013, Mar 26). Shale gas development impacts on surface water quality in Pennsylvania. Proceedings of the National Academy of Sciences of the United States of America, 110(13), 4962-4967. https://doi.org/10.1073/pnas.1213871110.

Rahm, B. G., Vedachalam, S., Bertoia, L. R., Mehta, D., Vanka, V. S., & Riha, S. J. (2015). Jul). Shale gas operator violations in the Marcellus and what they tell us about water resource risks. *Energy Policy, 82*, 1–11. https://doi.org/10.1016/j.enpol.2015.02.033

Roque, C., Lourenço Cardoso, J., Connell, T., Schermers, G., & Weber, R. (2019). 2019/10/01/). Topic analysis of Road safety inspections using latent dirichlet allocation: A case study of roadside safety in Irish main roads. *Accident Analysis & Prevention, 131*, 336–349. https://doi.org/10.1016/j.aap.2019.07.021.

Ryder, B., Gahr, B., Egolf, P., Dahlinger, A., & Wortmann, F. (2017). 2017/07/01/). Preventing traffic accidents with in-vehicle decision support systems - The impact of accident hotspot warnings on driver behaviour. *Decision Support Systems, 99*, 64–74. https://doi.org/10.1016/j.dss.2017.05.004.

Suh, Y. (2021). Oct). Sectoral patterns of accident process for occupational safety using narrative texts of OSHA database. *Safety Science, 142*, Article 105363. https://doi.org/10.1016/j.ssci.2021.105363

Torres, L., Yadav, O. P., & Khan, E. (2016). Jan 1). A review on risk assessment techniques for hydraulic fracturing water and produced water management implemented in onshore unconventional oil and gas production. *Science of the Total Environment, 539*, 478–493. https://doi.org/10.1016/j.scitotenv.2015.09.030

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). 2017/6//). *Attention Is All You Need.* http://arxiv.org/abs/1706.03762.

Wang, Y., & Xu, W. (2018). 2018/01/01/). Leveraging deep learning with LDA-based text analytics to detect automobile insurance fraud. *Decision Support Systems, 105*, 87–95. https://doi.org/10.1016/j.dss.2017.11.001.

Ye, M., Osman, O. A., Ishak, S., & Hashemi, B. (2017). 2017/09/01/). Detection of driver engagement in secondary tasks from observed naturalistic driving behavior. *Accident Analysis & Prevention, 106*, 385–391. https://doi.org/10.1016/j.aap.2017.07.010.

Zhu, G., Wang, Y., Cao, J., Bu, Z., Yang, S., Liang, W., & Liu, J. (2021, 2021/01/09/). Neural Attentive Travel package Recommendation via exploiting long-term and short-term behaviors. Knowledge-Based Systems, 211, 106511. https://doi.org/https://doi.org/10.1016/j.knosys.2020.106511.

Dan BI is currently pursuing the PhD degree in management science and engineering at School of Management, Xi'an Jiaotong University, Xi'an, China. Her research interests include unconventional oil and gas environmental issue, artificial intelligence, nature language processing.