Identify Environmental And Health Threats In Unconventional Oil And Gas Violations: Evidence From Pennsylvania Compliance Reports

Dan Bi

Xi'an Jiaotong University

Ju-e Guo

Xi'an Jiaotong University

Erlong Zhao

Xi'an Jiaotong University

Shaolong Sun (✉ sunshl13@lzu.edu.cn)

Xi'an Jiaotong University <https://orcid.org/0000-0002-3196-1459>

Shouyang Wang

Chinese Academy of Sciences

Research Article

Keywords: Unconventional oil and gas, text mining, LDA, environmental violation, decision tree, environment pollution

Posted Date: September 7th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-760912/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

With unconventional oil and gas booming commercial development, its inevitable damage to the environment arouses the vigilance of the public. It is of great necessity both for practitioners and governors, to learn from the history, find the pattern of violations then build a sound production warning system. In this respect, we utilized the “Oil and Gas Compliance Report” from the Pennsylvania Department of Environmental Protection from 2000 to 2019. After the preprocessing, a total of 5737 violation records were obtained. Through LDA (Latent Dirichlet Allocation) analysis combined with the decision tree model, our research attained the following conclusions: first, we find that the LDA themes of violations as “Erosion and sediment” and “Water pollution” are key factors for “Failed” enforcement results. Therefore, policymakers and practitioners should pay more attention to these two types of accidents. Second, it is noted that counties are also one of the important features that matter the enforcement results. Third, we need to consider the role of economic punishment dialectically, while it is not a significant feature for successful enforcement results. That is to say, monetary punishment may not necessarily improve the effectiveness of the company’s handling of violations.

Introduction

With unconventional oil and gas booming commercial development, its inevitable damage to the environment arouses the vigilance of the public. There are multiple types of violations that threaten human health and environmental safety caused during its production process, including soil erosion, groundwater pollution, geological damage, and so on. It is of great necessity both for practitioners and governors, to learn from the history, find the pattern of violations then build a sound production warning system. In particular, the dataset of unconventional oil and gas records in compliance reports, the narrative text of specific items named “violation code description” is of great importance to analyze the cause, resource, release substances, and misbehaviors during accidents. In this respect, this study aims to dig out the pattern and themes of violations, using the Latent Dirichlet Allocation (LDA) and decision tree to draw a roadmap that happened in the last 20 years.

Although there is an enormous amount of research about unconventional oil and gas development in recent years, there is still a missing gap for environmental violations causal analysis. Rivard et al. (2014) introduced the current status of Canada’s rising unconventional oil and gas industry, also emphasized public concerns about hydraulic fracturing technology and the environmental problems it causes, which is particularly prominent in Quebec. Krupnick et al. (2014) made a macro-level assessment of the environmental risks of China’s unconventional oil and gas development, drawing on the opinions of the U.S. unconventional oil and gas development, starting from environmental regulations and environmental assessment benchmarks—water, air, and land, and pointed out that there is still lack of unconventional oil and gas strategic environmental assessment and industry environmental regulations in China. At present, most articles on the establishment of pollution indicators are based on expert opinions and lack objective model explanations and reliable data support.

Natural language processing is an intersection of computer science and artificial intelligence. It aims to develop artificial intelligence algorithms that can effectively realize natural language communication, which plays a significant support role during the human decision-making process (Chen et al., 2019; Demoulin and Coussement, 2020; Ding et al., 2021). Using natural language processing and artificial intelligence algorithms to dig out more decision-making information (Li et al., 2020; Jung and Lee, 2020), such applications are widely used in the field of information management research. For example, online shopping reviews analysis (Li et al., 2020), hotel and restaurant reviews analysis (Jia, 2020; Nicolau et al., 2020), and even the voluntary national reviews released by the United Nations (Sebestyén et al., 2020), etc. This technology transforms unstructured data into structured data, enriching the original data, combined with artificial intelligence methods can accomplish much more intriguing research (Feuerriegel and Gordon, 2018; Carracedo et al., 2020; Ficcadenti et al., 2019; Guerreiro and Rita, 2020).

The purpose of this article is to combine text mining and machine learning to explore what affects the consequences of violations enforcement. We used the “Oil and Gas Compliance Report” from the Pennsylvania Environmental Protection Agency. The period is from 2000 to 2020, after the preprocessing, 5737 violation records were obtained. Through LDA analysis combined with the decision tree model, our research aims to solve the following two questions:

- What themes can be obtained from historical environmental violation records?
- What pattern can be obtained and used to alarm for this industry development?

The structure of this paper is arranged as follows: **Sect. 2** introduces literature about the environmental issues and regulations of unconventional oil and gas; Sect. 3 makes a statistical analysis of unconventional oil and gas violations; Sect. 4 presents the framework and methods; then Sect. 5 conducts LDA analysis and displays the themes of regulations; we combine LDA document vectors with other dummy variables to establish a decision tree model in Sect. 6; in Sect. 7, we provide the significance and future research.

1.1 Study area and data source

The U.S. natural gas production in 2019 was approximately 34 trillion cubic feet. Among the 35 natural gas-producing states, five states accounted for approximately 69% of the total U.S. dry natural gas in 2019. They are: Texas (23.9%), Pennsylvania (20 %), Louisiana (9.3%), Oklahoma (8.5%) and Ohio (7.7%) (EIA). Among them, to provide information, the Pennsylvania State Department of Environmental Protection has established an oil and gas report database in a user-friendly format, the recorded content is reliable and instructive (Abualfaraj et al., 2016; Kim and Oliver, 2017; Krupnick et al., 2014). In this paper, we collected data from the official Pennsylvania government website, the Department of Environmental Protection “Oil and Gas Compliance Report”, then choose the oil type as “unconventional oil and gas”.

Literature Review

Shale gas and hydraulic fracturing have revolutionized the US energy sector in terms of prices, consumption, and CO₂ emissions. However, the key questions remaining include environmental concerns and extraction efficiencies that are leveling off (Middleton et al., 2017). The points in the dispute over environmental issues can be divided into three categories: air, water, and land. Air pollution is mainly reflected in greenhouse gas emissions. At present, there is still controversy about the effectiveness of greenhouse gas emission reduction of this industry; the impact on water is reflected in a large amount of freshwater demand for its production and the risk of water pollution during fracturing; while the hydraulic fracturing and drilling will affect the local geological structure, landform, biodiversity, etc.(Jenner and Lamadrid, 2013).

The worldwide development of shale gas has allowed scholars to pay attention to the different risks associated with the unique geological conditions of different countries. Krupnick et al. (2014) summarized the issues and activities of concern relative to the shale and tight gas industry, also emphasized public concerns about hydraulic fracturing in Canada, Quebec. Prpich et al. (2016) and Tagliaferri et al. (2017) updated the environmental risks related to shale gas development in the UK and discussed how operators and regulators can cooperate under a high-level environmental risk assessment framework. The potential impact on water resources of the illegal accidents in the Bakken shale gas area in Canada has been studied, and Shrestha et al. (2017) have proposed measures that may be used to minimize water resources risks. Through a questionnaire survey of 730 residents living in shale gas-producing areas in China's Sichuan Basin, the residents' perceptions of shale gas development risks were discussed in this work(Yu et al., 2018). They also pointed out that there are still vacancies in China's shale gas mining regulations, and there is no independent, comprehensive, and strict regulatory system to solve potential environmental risks.

Many scholars have set their sights on the research of environmental pollution issues during unconventional oil and gas development. Manda et al. (2014) used geospatial and statistical analysis of 3786 environmental violation accidents to reveal the impact of the single-well pad and multi-well pad in shale gas on wastewater management and environmental pollution. Logistic regression was used to test the probability of violations of unconventional and conventional oil under different conditions, as well as the differences in the probability of violations between different development operators were found, and more complicated efforts should be made to explain the performance of each company (Abualfaraj et al., 2016). Kim and Oliver (2017) used shale gas inspection reports from 2000 to 2013 to explore the effectiveness of Pennsylvania's newly revised Oil and Gas Act and found that the current environmental regulations have indeed played a role in reducing violations. Guo et al. (2019) used the relationship between crime and punishment as an analysis framework to discuss environmental violations of shale gas in Pennsylvania. The study found that the type of inspection activities and the degree of economic punishment are important factors that affect whether environmental accidents are violated.

Compared with conventional energy coal, unconventional oil and gas have both cost and profit advantages in residential, commercial, and industrial utilization. At the same time, it is also the cleanest fossil energy (Nature, 2009). Therefore, the most imperative issue of this industrial development is to

strengthen and powerfully implement the environmental regulations with an appropriate sustainable goal, which is the best way to balance the development of both economic and environmental. Kuwayama et al. (2017) conducted a statistical analysis of violation data in New Mexico, Pennsylvania, Colorado, and Oklahoma, then provided a critical review of environmental regulations for wastewater and on-site storage.

To improve the establishment of unconventional oil and gas environmental regulations, how to determine the path of environmental pollution and assess the level of environmental pollution is an important question in research. We believe that the research paradigm is undergoing a path from expert experience to quantitative research based on statistics. Krupnick et al. (2014) used the American experience to combine with China's geological conditions and environmental regulations, they further discussed the potential risks of China shale gas development and emphasized the importance of strategic environmental assessment for future sustainable development. Krupnick and Gordon (2015) had conducted a questionnaire survey on 215 experts in the shale gas industry and summarized a comprehensive range of possible environmental pollution and risk paths of shale gas, which is also the first approach based on survey and statistical analysis. Rahm et al. (2015) conducted a statistical-based study on 3267 unconventional oil and gas violations released by the Pennsylvania Environmental Protection Agency and found that environmental policies are highly correlated with environmental violation rates. The study concluded that leakage and solid waste management involved in shale gas development are riskier. However, most studies have pointed out the need for further experiments due to the lack of field data (Krupnick et al., 2014).

Framework And Methodologies

The framework of this article is shown in Fig. 1. In the first step, we obtain the topic distribution of environmental regulation through text mining. The period is from 2000 to 2019, a total of 5737 compliance reports with "Enforcement final status" are applied. To further elucidate the complex relationship of regulation topics, some dummy variables and "Enforcement final status" are all applied in the decision tree model.

Figure 2 shows the proportional distribution of the enforcement final status. The reason why we use "enforcement final status" as the dependent variable is that if the consequences of violations are successful, that is, "Comply/Closed", then the harm to the environment due to the accident can be minimized. It also reflects the successful efforts of both the enterprise and the government. When the consequences of a violation accident are "Failed", there are maybe two reasons behind: first, the scope of this accident is huge and cannot be solved even though the great efforts of both the government and enterprises; second, the consequences of accidents on the environment are irreversible, and the damage to the environment cannot be restored or reduced. Therefore, the purpose of a decision tree model is to obtain important features that affect the enforcement results from historical data and provide a solid support for supervision and environmental regulations.

In the process of constructing text topic and visualization, this article uses the text topic model LDA and t-SNE. The purpose of the former is to form different themes and their corresponding document-topic matrices, the latter is to reduce the dimensionality of the high-dimensional document-topic matrices and present them in a 2-dimensional form to show the similarities and differences of themes more intuitively.

LDA:

LDA (Latent Dirichlet Allocation) is a document topic generation model, also known as a three-layer Bayesian probability model, which contains a three-layer structure of words, topics, and documents. The so-called generative model means that each word in an article is obtained through the process of "selecting a certain topic with a certain probability, and selecting a certain word from this topic with a certain probability". Documents to topics follow a polynomial distribution, and topics to words follow a polynomial distribution. LDA is an unsupervised machine learning technology that can be used to identify the subject information hidden in a large-scale document collection or corpus.

t-SNE:

t-Distributed Stochastic Neighbor Embedding (t-SNE) is a dimensionality reduction technique used to represent high-dimensional data sets in a low-dimensional space to visualize them. Compared with other dimensionality reduction algorithms (such as PCA), t-SNE creates a reduced feature space, similar samples are modeled by nearby points, and dissimilar samples are modeled by high-probability distant points. At a high level, t-SNE constructs a probability distribution for high-dimensional samples. Similar samples are highly likely to be selected, while different points are very unlikely to be selected. Then, t-SNE defines a similar distribution for the points in the low-dimensional embedding. Finally, t-SNE minimizes the Kullback-Leibler (KL) divergence between the two distributions about the position of the embedding point.

Decision tree:

Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression problems. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. In this study, we use the Gini index to choose partition attributes, which is based on the Classification and Regression Tree (CART) decision tree algorithm (Breiman et.al., 1984). Supposing that the proportion of the k th attributes in the current sample set D is p_k ($k = 1, 2, \dots, |y|$), the purity of the data set D can be measured by the Gini value:

$$Gini(D) = \sum_{k=1}^{|y|} \sum_{k' \neq k} p_k p_{k'} = \sum_{k=1}^{|y|} p_k^2 \quad (1)$$

Intuitively speaking, $Gini(D)$ reflects the probability that two samples are randomly selected from the data set D and their category labels are inconsistent. Therefore, the smaller the $Gini(D)$, the higher the purity of

the data set D . In the candidate attribute set A , the attribute with the smallest Gini value after division is selected as the optimal division attribute.

3.1 Experiment setup

We utilize python 3.7.x as the programming environment. The API of NLTK 3.6.2 version and GENSIM 3.8.3 is applied in narrative text analysis. The algorithm of the decision tree is from Scikit-learn 0.22. We first preprocessed the text, including tokenization, stopwords removal, and word lemmatization. The original material contains 5737 documents, after processing, 783 unique tokens are obtained. After that, we brought the LDA model to obtain regulatory topics from 2000 to 2019. In this study, we determined the number of topics is 5 acquired the best performance of LDA.

We use the cost complexity pruning to control the size of the tree instead of parameters such as minimum samples leaf and maximum depth. Minimal cost-complexity pruning is an algorithm used to prune a tree to avoid over-fitting (L. Breiman et al., 1984) This algorithm is parameterized by $\alpha \geq 0$ known as the complexity parameter. Minimal cost complexity pruning recursively finds the node with the “weakest link”. The weakest link is characterized by an effective α , where the nodes with the smallest effective α are pruned first. As α increases, more of the tree is pruned, thus creating a decision tree that generalizes better. We set $\alpha = 0.00047$ to maximize the testing accuracy, which reaches 89%.

Data Description

The data period of this study is 19 years, from 2000 to 2019, the whole number of violations is 10,440, after preprocessing, we attained a total of 5737 violation records with enforcement final status. The objects that we focus on and their corresponding description are listed in Table 1, detailed statistical data is listed in Appendix A, from Table A1 to Table A8.

Table 1
Data description.

Objects	Description
Inspection type	There is a total of 20 different types of inspections, which are: "Routine/Complete Inspection", "Drilling/Alteration", "Routine/Partial Inspection", "Follow-up Inspection", "Incident- Response to Accident or Event", "Site Restoration", "Administrative/File Review", "Compliance Evaluation", "Complaint Inspection", "Preoperation Inspection", "Plugging(Includes Plugged/Mined Through)", "Site Restoration Extension", "Routine/Partial/Aerial Inspection", "Bond Release", "Construction", "Routine Final Inspection", "Compliance Schedule Evaluation", "Facility Operations Inspection", "Asbestos Program Inspection", "Joint Internal Site Inspection".
County	The county of Pennsylvania where the facility is located.
Inspection result description	The description of the associated inspection result, which are: "Violation(s) noted", "Outstanding violations", "Violations noted & Immediately Corrected", "Violations & outstanding violations", "Violations & recurring violations", "Recurring violations".
Violation date	The date the violation was identified, which from 2000/1/1 to 2019/12/31.
Violation code & description	The code and description of the violation that was identified during the inspection. Such as 102.11 - Failure to design, implement, or maintain BMPs to minimize the potential for accelerated erosion and sedimentation.
Violation type	The type of violation (Administrative or Environmental Health & Safety).
Enforcement code & description	The codes and description associated with the enforcement, which are: NOV - Notice of Violation CACP - Consent Assessment of Civil Penalty COA - Consent Order and Agreement ADORD - Administrative Order CMPOR - Compliance Order CESOR - Cessation Order EHBO - Environmental Hearing Board Order FLNOV - Field Notice of Violation CPA - Civil Penalty Assessment
Enforcement final status	The description is used to identify the final status of enforcement. There is a total of 6 types of final status, which are "Comply/Closed", "Administrative Close Out", "Superseded", "Unresolvable", "Vacated", "NOV Rescinded".

After a comprehensive analysis of the data, we found that the "Violation code" in the violation report largely explained the cause of this violation, while "Inspection result description" assisted in explaining the scale and severity of the violation. "Enforcement code description" is the administrative measure for

this violation, and “Enforcement final status” represents the result of the violation, whether it was successful remediation or other failures.

Violations Topic Model

In this section, we will introduce the steps for building violation regulations topic models and further interpret the generating topics.

There is a total of 5737 documents after preprocessing. The structure of each “violation code” is as: number of regulation chapters + abbreviation of regulation-specific violation content. For example, “201TAG-Failure to install, in a permanent manner, the permit number on a completed well”. This article only selects the specific content after “-” to analyze the theme of violation accidents.

The themes of all violations and “Environmental Health & Safety” are both divided into five categories, their themes are highly similar. The first theme is “Residual waste management”, this type of environmental regulation usually addresses the emergency treatment of pollutant leakage during the production process, such as vehicle diesel leakage during transportation or material leakage during wastewater management. The second topic is “Drilling fluid spill”. This type of environmental regulation is aimed at environmental violation accidents caused by hydraulic fracturing, usually related to the improper discharge of return water, or sewage leakage during fracturing. The third theme is “Emergency response”, in this case, operators delayed response and report to the corresponding agency, which results in the scale of pollution accidents to expand and cause severe environmental damage. While the fourth topic is “Erosion and sediment control”, a type of violation of regulations aimed at situations where on-site work fails to minimize accelerated erosion and stabilize the site until restoration. The fifth theme is “Water pollution”, which mainly contains two situations: discharge of pollution material to waters of the commonwealth, operator failed to prevent gas and other fluids from lower formations from entering fresh groundwater.

There are two themes for “Administrative” violation codes. The first theme is “Emergency report”, in most cases, the operator failed to post the permit number, operator name, address, telephone number in a conspicuous manner. The second theme is “Well integrity”, which means operators failed to construct and operate the well to ensure that the integrity of the well is maintained.

Table 2
Themes of violations' regulation.

Corpus	# Topic	Themes	Top 10 relative tokens
All	1	Residual waste management.	Failure, waste, residual, dispose, properly, permit, water, commonwealth, pollution, process.
	2	Drilling fluid spill.	Well, failure, drilling, fluid, pollution, within, operator, completion, failed, permit.
	3	Emergency response.	Dep, waste, person, residual, facility, regulation, plan, emergency, code, maintain
	4	Erosion and sediment.	Site, used, failure, plan, implement, specific, general, emergency, code, maintain.
	5	Water pollution.	Operator, water, commonwealth, failed, well, tank, material, discharge, polluttional, fluid
Environmental Health & Safety	1	Water pollution.	Permit, water, commonwealth, issued, discharge, potential, polluttional, material, polluting, without.
	2	Residual waste management.	Waste, residual, failure, dispose, fluid, properly, water, store, process, transport.
	3	Drilling fluid spill.	Well, drilling, fluid, equipment, tank, failure, operator, pit, failed, produced.
	4	Emergency response.	Operator, well, failed, response, within, emergency, planning, safety, site, health.
	5	Erosion and sediment.	Failure, site, plan, emergency, pollution, implement, maintain, erosion, response, control.
Administrative	1	Emergency report	Failure, used, well, code, general, specific, cannot, impoundment, number, act.
	2	Well integrity.	Failure, pit, plan, well, permit, drilling, oil, gas, substance, construct.

We also note that the types of violation accidents can be divided into two categories, one is "Environmental health & safety", and the other is "Administrative". To reveal the subtle differences of these two types of violations, we respectively utilize 3451 "environmental" violation codes and 2286 "Administrative" violations codes to build LDA models, and the optimal number of topics was 5 and 2. To better show the effect of topic classification, we have done t-SNE dimensionality reduction based on the document-topic matrix of LDA. The themes of LDA documents are shown in Table 2, and the effect of LDA classification is shown in Figures 4 and 5, and each dot represents a document.

t-SNE is generally used as a tool for dimensionality reduction and extraction of main elements, which can further explain the correlation between different themes. In Fig. 3, which is based on the whole corpus, topic 1, 2, and 5 can be well separated from others, while the boundary between topic 3 and topic 4 is not

obvious, even some areas overlap after dimensionality reduction. It implicates the high correlation between topic 3 and topic 4, that is, “Emergency response” and “Erosion and sediment” usually appear together in the violation regulations. As for Fig. 4, topic 5 and topic 2 are well distinguished from other topics, but topics 1, 3, and 4 overlap each other. This means that topics 5 and 2 usually appear as separate topics among violation regulations, but the content between the topics “Water pollution”, “Drilling fluid spill”, and “Emergency response” may complement each other and appear together in the violation regulations.

Decision Tree Results And Discussion

In this section, we use the previously mentioned independent factors in Table 2 to build a decision tree model. The dependent variable of the decision is: “Enforcement final status”, which consists of two conditions: “Comply/Closed” and “Failed”. As a white box model, the decision tree can explain the degree of influence of each independent variable on the dependent variable.

Except for the “Violation code”, which is presented in the form theme-document matrix, other independent variables are presented as dummy variables, that is, 1 means occurrence and 0 means no. The overall variables of the decision tree are shown in Fig. 5 below:

Table 3 is the top 35 feature that have the impact for enforcement results of violations. It is noted that “Violation date”, “County”, “Enforcement code description”, themes of violations obtained by LDA, “Inspection type”, “Inspection result description”, and “Inspection source” are the top 7 important impactors. But the relationship between those factors can’t visualize through Table 3, we will use decision tree roadmap to show the patterns between bellow factors.

Table 3
Key factors for enforcement results

Rank	Feature name	Feature importance
1	Violation date	0.191363
2	County Susquehanna	0.121717
3	Enforcement code: CACP - Consent Assessment of Civil Penalty	0.089137
4	Topic 4	0.075278
5	County Somerset	0.059117
6	County Greene	0.047737
7	County Washington	0.046644
8	Enforcement code: NOV - Notice of Violation	0.037634
9	Topic 2	0.035327
10	Topic 3	0.034699
11	County Wyoming	0.023694
12	Topic 5	0.022939
13	Inspection type: Incident- Response to Accident or Event	0.019401
14	Inspection source: EFACTS	0.01854
15	Inspection result: Violation(s) & Outstanding Violations	0.016991
16	County Sullivan	0.014569
17	County Armstrong	0.012632
18	County Westmoreland	0.012374
19	County Bradford	0.011777
20	Enforcement code: ADORD - Administrative Order	0.010974
21	Topic 1	0.010586
22	Inspection type Compliance Evaluation	0.009899
23	Inspection type Follow-up Inspection	0.009394
24	Inspection type Routine/Partial Inspection	0.009326
25	Violation type Administrative	0.009228
26	Inspection source: SUBSAIR	0.008389

Rank	Feature name	Feature importance
27	Enforcement code description: COA - Consent Order and Agreement	0.006555
28	Inspection type: Routine/Complete Inspection	0.006465
29	County Lycoming	0.005136
30	Enforcement code description: EHBO Environmental Hearing Board Order	0.005134
31	Inspection type: drilling/alteration	0.004463
32	County Huntingdon	0.003708
33	County Tioga	0.003408
34	Inspection type: Complaint Inspection	0.003225
35	County Clearfield	0.002538

The decision tree constitutes an “if-then” conditional logic sequence, and its interpretation sequence is from top to bottom (Fig. 6). The decision degree depth after pruning has 8 layers of depth, the 0th level is the root node (depth 0, top), and the corresponding influencing factor of the root node is “County Susquehanna”. Assuming that there is a training sample currently, the root node asks whether the county is Susquehanna (Boolean variable), if the value is greater than 0.5, then false, move down to the left side of the root node (depth 1, left 1), in this case, the decision tree continues to ask whether the violation accident occurred less than 2011, if true, move down to the left node (depth 2, left 1). In this case, it is a leaf node, then we can check the prediction class of the node, the decision tree predicts that the violation accident can be resolved (class: Comply/Closed, Gini: 0.143).

The “sample” attribute of each node calculates the number of training examples. For example, the leaf node (depth 1, left 1), 4063 training samples satisfy the county is not Susquehanna, 526 training samples satisfy the county is Susquehanna (depth 1, left 2). The “value” attribute of the node indicates the number of training instances of each class used by this node: for example, the root node (depth 0, top) is suitable for 3627 cases of Comply/closed, and 962 cases of Failed. Finally, the “Gini” attribute of a node measures its impurity: if all training examples it uses belong to the same class, the node is “pure” (Gini = 0).

The colors of the decision tree nodes represent different classes. The purple nodes represent the class “Comply/Closed”, and the yellow represents “Failed”. The more obvious the color, the better the classification effect (the higher the purity). From the pruned decision tree, there are some enlightening correlations between the impactors and violation final status. When the result of the violation is a failure, there are two judgment chains with significantly small Gini coefficient: violation date is over 2012 → enforcement code description is “Notice of violation” → county is Somerset → then violation final status is “Failed”; while the second one is: Violation date is over 2012 → Enforcement code description is “Notice

of violation" → county is Washington → the violation LDA topic is "Erosion and Sediment" → the violation final status is "Failed".

There are two judgment chains for successfully handling violations with a lower Gini coefficient (higher purity), namely: the county is not Susquehanna → violation date is over 2011 → enforcement code is not "Notice of Violation" → the violation final status is "Comply /Closed"; while the second is that county is Susquehanna → enforcement code is "Consent Assessment of Civil Penalty" → then the violation final status is "Comply/Closed".

From the above judgment chains we can obtain the following implications for policymakers and practitioners:

First of all, the LDA classification of violation code helps policymakers to pay more attention to related environmental regulations. In the pruned decision tree, the LDA topics as node features are Topic 4 "Erosion and Sediment" and Topic 5 "Water Pollution". When the violations involve the above two themes, most of the consequences of the accident (more than 50%) are unsuccessful.

Second, economic punishment is not a significant feature that enables the successful enforcement of violations. From the perspective of higher purity sub-nodes, factors like: "Enforcement code", "Violation date", "County", and LDA topics are highly connected with successful enforcement of a violation.

The enforcement consequences of violations in different counties are different, and our research has also pointed out counties with obvious differences. The counties that are of the high important feature in pruning decision trees include Susquehanna, Somerset, Greene, and Washington. Susquehanna had a significantly higher success rate for handling violations under the conditions of financial punishment (Gini = 0.122), and this county had a higher success rate for handling violations before 2008. Most of the violations in Greene and Somerset County after 2012 were not handled successfully. After 2012, besides the type of violation as "Erosion and sediment", the enforcement of violations generally failed in Washington (Gini = 0.165).

Conclusions And Policy Implications

The purpose of this study is to combine text mining and machine learning to dig out the main factors that affect the enforcement results of violations from the compliance reports. We hope to find meaningful factual relationships to remind policymakers and practitioners.

The work of this study can be divided into two steps. In the first step, we classify the "Violation code" into 5 categories of themes by LDA. After that, we combine the document-LDA topic matrix with other category variables as independent variables, while the "Enforcement final status" as dependent variables, bring them into the decision tree model. In the end, we obtain the causal logic chains and important features that cause serious violations.

First, we find that the LDA themes of violations as “Erosion and sediment” and “Water pollution” are key factors for “Failed” enforcement results. Therefore, policymakers and practitioners should pay more attention to these two types of accidents. Second, it is noted that counties are also one of the important features that affect the classification of the decision tree. Among them: Susquehanna and Greene, most of their violations after 2012 have not been successfully handled. Third, we need to look at the role of economic punishment dialectically, while it is not a significant feature of successful handling of violations as its Gini coefficient is only 0.5, which means that more than half of the accidents still have economic punishments, but the processing results are still failures.

The innovation of this study is that text mining technology is firstly applied in the identification of impactors from unconventional oil and gas violations, the database for unconventional oil and gas violations analysis has been expanded, which can provide evidence for establishing environmental pollution detection indicators; besides, this study makes full use of the unstructured data in the unconventional oil and gas reports, combined with the decision tree algorithm to distinguish the important characteristics and logic chain of unconventional oil and gas violations.

The contribution of this study lies in two aspects. First, it provides a new perspective on unconventional oil and gas environmental violations. We use natural language processing to combine unstructured text data that cannot be quantified and almost be omitted in previous studies with a white box machine learning approach, which greatly improves interpretability. Second, this study uses the LDA classification to avoid the preconceived classification subjectivity of previous studies and objectively extracts different topics contained in the original violation corpus, which can greatly assist policymakers to identify different aspects of environmental accidents, combined with local circumstances to find the characteristics and distinctive pattern.

One of the limitations of this study is that the LDA model cannot extract complex logical and semantic meanings in the content. Therefore, we reckon that our future research is to develop a task-oriented deep learning model to understand the relationship between violations and environmental regulations; although our research has pointed out the special performance of certain counties, due to the lack of relevant information, it is impossible to analyze the reasons behind it.

Declarations

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>].

Competing interests

The authors declare that there are no conflicts of interest regarding the publication of this study.

Funding

This research work was partly supported by the National Natural Science Foundation of China under Grant Nos. 71774130 and 71988101, and the Fundamental Research Funds for the Central Universities under Grant No. xpt012020022.

Authors' contributions

All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Dan Bi and Erlong Zhao. The first draft of the manuscript was written by Dan Bi and all authors commented on previous versions of the manuscript. Ju-e Guo supervised analysis process. Shaolong Sun, Ju-e Guo, and Shouyang Wang acquired the funding support All authors read and approved the final manuscript.

References

1. Abualfaraj N, Olson MS, Gurian PL, De Roos A, Gross-Davis CA, 2016, Statistical analysis of compliance violations for natural gas wells in Pennsylvania: *Energy Policy*, v. 97, p. 421–428
2. Buono RM, Mayor B, López-Gunn E (2018) A comparative study of water-related issues in the context of hydraulic fracturing in Texas and Spain, v. 90. *Environmental Science & Policy*, pp 193–200
3. Carracedo P, Puertas Medina R, Luisa Martí M Selva (2020) Research lines on the impact of the COVID-19 pandemic on business. *Journal of Business Research*, A text mining analysis
4. Chen Y, Zhang H, Liu R, Ye Z, Lin J, 2019, Experimental explorations on short text topic mining between LDA and NMF based Schemes: *Knowledge-Based Systems*, v. 163, p. 1–13
5. Cotton M, Rattle I, Van Alstine J, 2014, Shale gas policy in the United Kingdom: An argumentative discourse analysis: *Energy Policy*, v. 73, p. 427–438
6. Demoulin NTM, Coussement K, 2020, Acceptance of text-mining systems: The signaling role of information quality: *Information & Management*, v. 57, p. 103120

7. Ding Z, Liu R, Yuan H (2021) A text mining-based thematic model for analyzing construction and demolition waste management studies. *Environmental Science Pollution Research* 28:30499–30527
8. Feuerriegel S, Gordon J, 2018, Long-term stock index forecasting based on text mining of regulatory disclosures: *Decision Support Systems*, v. 112, p. 88–97
9. Ficcadenti V, Cerqueti R, Ausloos M, 2019, A joint text mining-rank size investigation of the rhetoric structures of the US Presidents' speeches: *Expert Systems with Applications*, v. 123, p. 127–142
10. Guerreiro J, Rita P (2020) How to predict explicit recommendations in online reviews using text mining and sentiment analysis. *Journal of Hospitality Tourism Management* v 43:269–272
11. Guo M, Xu Y, Chen YD, 2019, Environmental enforcement and compliance in Pennsylvania's Marcellus shale gas development: *Resources, Conservation and Recycling*, v. 144, p. 24–31
12. Hannibal B, Portney K (2020) The impact of water scarcity on support for hydraulic fracturing regulation: A water-energy nexus study. *Energy Policy* 146:111718
13. Jenner S, Lamadrid AJ (2013) Shale gas vs. coal: Policy implications from environmental impact comparisons of shale gas, conventional gas, and coal on air, water, and land in the United States. *Energy Policy* v 53:442–453
14. Jia SS (2020) Motivation and satisfaction of Chinese and U.S. tourists in restaurants: A cross-cultural text mining of online reviews. *Tourism Management* v 78:104071
15. Jung H, Lee BG, 2020, Research trends in text mining: Semantic network and main path analysis of selected journals: *Expert Systems with Applications*, v. 162, p. 113851
16. Kim B, Oliver ME, 2017, Taming drillers through legislative action: Evidence from Pennsylvania's shale gas industry: *Resource and Energy Economics*, v. 50, p. 15–35
17. Krupnick AJ, Gordon HG, 2015, What Experts Say About the Environmental Risks of Shale Gas Development: *Agricultural and Resource Economics Review*, v. 44, p. 106–119
18. Krupnick A, Wang Z, Wang Y, 2014, Environmental risks of shale gas development in China: *Energy Policy*, v. 75, p. 117–125
19. Kuwayama Y, Roeshot S, Krupnick A, Richardson N, Mares J (2017) Risks and mitigation options for on-site storage of wastewater from shale gas and tight oil development. *Energy Policy* v 101:582–593
20. Breiman L, Friedman J, Olshen R (1984) and C. Stone. *Classification and Regression Trees*. Wadsworth, Belmont
21. Li H, Liu H, Zhang Z (2020) Online persuasion of review emotional intensity: A text mining analysis of restaurant reviews. *International Journal of Hospitality Management* v 89:102558
22. Li Q, Zeng DD, Xu DJ, Liu R, Yao R, 2020, Understanding and Predicting Users' Rating Behavior: A Cognitive Perspective: *INFORMS Journal on Computing*, v. 32, p. 996–1011
23. Livy MR, Gopalakrishnan S, Klaiber HA, Roe BE (2018) The impact of intensity on perceived risk from unconventional shale gas development. *Journal of Environmental Management* v 218:630–638

24. Middleton RS, Gupta R, Hyman JD, Viswanathan HS, 2017, The shale gas revolution: Barriers, sustainability, and emerging opportunities: *Applied Energy*, v. 199, p. 88–95
25. Nicolau JL, Sharma A, Shin S, 2020, The tourism effect of President Trump's participation on Twitter: *Tourism Management*, v. 81, p. 104133
26. Prpich G, Coulon F, Anthony EJ, 2016, Review of the scientific evidence to support environmental risk assessment of shale gas development in the UK: *Science of The Total Environment*, v. 563–564, p. 731–740
27. Rahm BG, Vedachalam S, Bertoia LR, Mehta D, Vanka VS, Riha SJ, 2015, Shale gas operator violations in the Marcellus and what they tell us about water resource risks: *Energy Policy*, v. 82, p. 1–11
28. Rahm BG, Riha SJ (2012) Toward strategic management of shale gas development: Regional, collective impacts on water resources, v. 17. *Environmental Science & Policy*, pp 12–23
29. Rivard C, Lavoie D, Lefebvre R, Séjourné S, Lamontagne C, Duchesne M (2014) An overview of Canadian shale gas production and environmental concerns. *International Journal of Coal Geology* v 126:64–76
30. Sebestyén V, Domokos E, Abonyi J (2020) Focal points for sustainable development strategies—Text mining-based comparative analysis of voluntary national reviews. *Journal of Environmental Management* v 263:110414
31. Shrestha N, Chilkoor G, Wilder J, Gadhamshetty V, Stone JJ (2017) Potential water resource impacts of hydraulic fracturing from unconventional oil production in the Bakken shale. *Water Research* v 108:1–24
32. Sun J, Zhu X, Sun X (2020) Assessment and forecasting of eco-environmental early-warning system for shale gas production in a pressure-state-response framework: *Energy Sources, Part A. Recovery, Utilization, and Environmental Effects*, pp 1–20
33. Soares ALC, Pinto CC, Oliveira SC (2020) Impacts of anthropogenic activities and calculation of the relative risk of violating surface water quality standards established by environmental legislation: a case study from the Piracicaba and Paraopeba river basins, Brazil. *Environmental Science Pollution Research* 27:14085–14099
34. Tagliaferri C, Clift R, Lettieri P, Chapman C, 2017, Shale gas: a life-cycle perspective for UK production: *The International Journal of Life Cycle Assessment*, v. 22, p. 919–937
35. Vesalon L, Crețan R, 2015, 'We are not the Wild West': anti-fracking protests in Romania: *Environmental politics*, v. 24, p. 288–307
36. Wang C, Wang F, Du H, Zhang X (2014) Is China really ready for shale gas revolution—Re-evaluating shale gas challenges, v. 39. *Environmental Science & Policy*, pp 49–55
37. Yu C, Huang S, Qin P, Chen X (2018) Local residents' risk perceptions in response to shale gas exploitation: Evidence from China. *Energy Policy* v 113:123–134

Appendix

Figures

Figure 1

Framework of this study.

Figure 2

Enforcement final status.

Figure 3

Visualization of whole documents topics.

Figure 4

Visualization of “Environmental Health & Safety” documents topics.

Figure 5

Variables of the decision tree.

Figure 6

The pruned decision tree.