frontiers | Frontiers in Water

TYPE Original Research
PUBLISHED 17 August 2023
DOI 10.3389/frwa.2023.1123313

OPEN ACCESS

EDITED BY
Abbas Roozbahani,
Norwegian University of Life Sciences, Norway

REVIEWED BY
Lledó Castellet-Viciano,
University of Valencia, Spain
Hossein Khaleghian,
Oklahoma State University, United States

*CORRESPONDENCE
John C. Matthews
✉ matthews@latech.edu

RECEIVED 13 December 2022
ACCEPTED 03 August 2023
PUBLISHED 17 August 2023

CITATION
Betgeri SN, Vadyala SR, Matthews JC and Lu H (2023) Wastewater pipe defect rating model for pipe maintenance using natural language processing. *Front. Water* 5:1123313. doi: 10.3389/frwa.2023.1123313

Wastewater pipe defect rating model for pipe maintenance using natural language processing

Sai Nethra Betgeri[1], Shashank Reddy Vadyala[1], John C. Matthews[2]* and Hongfang Lu[3]

[1]Department of Computational Analysis and Modeling, Louisiana Tech University, Ruston, LA, United States, [2]Trenchless Technology Center, Louisiana Tech University, Ruston, LA, United States, [3]Southeast University, Nanjing, Jiangsu, China

Introduction: Closed-circuit video (CCTV) inspection has been the most popular technique for visually evaluating the interior status of pipelines in recent decades. Certified inspectors prepare the pipe repair document based on the CCTV inspection. The traditional manual method of assessing structural wastewater conditions from pipe repair documents takes a long time and is prone to human mistakes. The automatic identification of necessary texts has received little attention. Computer Vision based Machine Learning models failed to estimate structural damage because they are not entirely understood and have difficulty providing high data needs. Hence, they have problems providing physically consistent findings due to their high data needs. Currently, a very small curated annotated image and video data set with well-defined, precisely labeled categories to test Computer Vision based Machine Learning models.

Methods: This study provides a valuable method to determine the pipe defect rating of the pipe repair documents by developing an automated framework using Natural Language Processing (NLP) on very small, curated annotated images, video data, and more text data. The text used in this study is broken into grammatical units using NLP technologies. The next step in the analysis entails using words to find the frequency of pipe defects and then classify them into respective defect ratings for pipe maintenance.

Results and discussions: The proposed model achieved 95.0% accuracy, 94.9% recall, 95% specificity, 95.9% precision score, and 95.7% F1 score, showing the potential of the proposed model to be used in large-scale pipe repair documents for accurate and efficient pipeline failure detection to improve the quality of the pipeline.

KEYWORDS

defect detection, wastewater pipe inspection, natural language processing (NLP), text recognition, trenchless technology

1. Introduction

The underground pipeline system forms a significant part of the infrastructure because it includes thousands of miles in the United States. Sanitary wastewater collects wastewater from public and private users as part of wastewater infrastructure systems (Mohammadi et al., 2019; Moradi et al., 2020). About 500,000 miles of private wastewater laterals and 800,000 miles of municipal wastewater lines (Malek Mohammadi et al., 2019). By 2032, 56 million people are expected to use centralized treatment plants (Nicklow et al., 2010; Vladeanu and Matthews, 2019a,b; Betgeri and Smith, 2021; Betgeri et al., 2023b). Water supply and wastewater water pipelines are essential for society's survival, and their security and efficiency are critical for human

health and economic growth (Yugandhar and Nethra, 2014; Li et al., 2016, 2019; Cheng and Wang, 2018; Hassan et al., 2019; Vladeanu and Matthews, 2019a; Betgeri, 2022; Boskabadi et al., 2022; Betgeri et al., 2023a). Using risk-based asset management, the most critical assets to take the most efficient course of action are identified by prioritizing the highest risk of failure by considering all the parameters.

Using the traditional manual method, the number of failures received by wastewater management can increase rapidly, making pipe failure handling imperative because the inspectors manually produce them by checking through CCTV films, manually recognizing and classifying such failures through pipe repair paperwork, and extracting the information connected to those pipe failures is difficult. As a result, the manual extraction procedure has a high potential for human mistakes, time consumption, and information loss. This issue can be resolved by substituting autonomous computational extraction for these manual procedures. The difficulty of computationally extracting information from free-text narratives is addressed by the field of study known as information extraction (IE), a subfield of natural language processing (NLP).

NLP can be an efficient way of automatically extracting information from large-scale pipe repair documents. NLP is a computer-assisted approach for facilitating the processing of human (natural) language. NLP may be a research area developing techniques accustomed to analyzing and extracting valuable data from text and speech in natural languages. Several NLP applications include information extraction, language translation, and opinion mining (Cambria and White, 2014; Vadyala and Betgeri, 2021; Vadyala and Sherer, 2021; Vadyala et al., 2022a,b). Text classification using NLP has been used for many problems within pipeline construction: To classify construction documentation based on priority (Yugandhar and Nethra, 2014; Vadyala et al., 2021). To support field inspection and data extraction of the inspection (Zhong et al., 2019). Information from work hazard analyses is processed using an ontology-based text categorization approach (Caldas and Soibelman, 2003). NLP methods are often divided into two main categories: (1) Rule-based and (2) Machine learning (ML) (Chi et al., 2014). Systems that rely only on hand-coded syntactic rules are known as rule-based systems (Le and David Jeong, 2017). As a result, their performance is underwhelming. Languages and linguistic grammar are unimportant in the Machine Learning-based approach (Marcus, 1995) because patterns are often quickly learned from unclear training examples because they outperform the state-of-the-art model like *K*-NN.

Nevertheless, NLP is a valuable technique for extracting and processing information from natural language into a more organized format for study. Natural Language Toolkit (NLTK) systems may be automated to parse textual content and search for keywords and phrases to extract data using predefined computer algorithms. The following is how key phrase extraction may be expressed as a sequence labeling task. Predict a sequence of labels, one label for each word in the input, where each label key phrase word) or non-KP, given an input sequence x, where each x represents the input vector of the keyphrase word). The task formulation for sequence labeling considers the correlations between nearby labels. It enables the joint decoding of the optimal sequence of labels for the input sequence rather than decoding each label individually.

Inspired by advancements in natural language processing, some researchers have more recently applied recurrent neural networks (RNNs) to wastewater pipe assessment documents for extraction and classification (Cosham and Hopkins, 2004; Graves et al., 2013; Dang et al., 2018; Jallan, 2020; Chahinian et al., 2021). Recurrent neural networks (RNNs) include Long Short-Term Memory Networks (LSTMs) that solve the problem of RNNs' gradient disappearing. Additional memory cells in LSTMs are used to store memories from long-distance phrases. Because LSTMs may store information from past sequence inputs in the current input state, they have proven a natural option for data applications such as speech recognition, language modeling, and trial option (Niu and Srivastava, 2022). An LSTM has a hidden layer, an input layer, and an output layer (Endalie et al., 2022). The hidden state in a forward LSTM network only saves information from the past. With the regular LSTM, input flows either in backward or forward directions. In bidirectional, input flows in two directions, creating a Bi-LSTM different from the regular LSTM. A bi-directional LSTM network with a forward hidden layer and a backward hidden layer to capture information flow in both directions is utilized (Yildirim, 2018; Yang and Zhao, 2020; Shafiei Alavijeh et al., 2021). The first model learns the sequence of the input provided, and the second model learns the reverse of that sequence. Data in the model is unstructured data to extract information from both sides at the entity level. The nodes in the hidden layer are linked, which is how long-distance information is kept in the matrix weights.

A comparison is performed between an LSTM and Bi-LSTM in pipe defect rating applications. LSTM and Bi-LSTM used many-to-many configurations for flaw detection and localization on simulated ultrasonic A-scans of holes and cracks. In their case, each exhibited perfect performance in the outputs of an LSTM in a many-to-many configuration as input to a dense decision-making layer for defect extraction, and assigning defect ratings on wastewater pipe assessment documents has historically proven to be challenging. It is unknown how these models generate particular decisions of defect classification and rating assignment because it is tough to interpret these data-driven models and how the rating is assigned to each defect characteristic. In addition, these methods are trained on small, curated data, and their generalization ability on unseen data is often limited (Tscheikner-Gratl et al., 2019; Wang, 2021). This paper proposed an ontology-based framework to improve efficiency and support decision-making regarding extraction and assign defect rating by automating the text classification by considering a complete set of defect Lexicon. The proposed pipe defect rating model uses the deep representation of entities using a knowledge base to reduce human efforts for labeled data creation and feature engineering. To illustrate the effectiveness of the proposed model, empirical experiments are conducted on a real dataset from the Department of Engineering and Environmental Services in Shreveport, Louisiana.

TABLE 1 Description of pipe inspection document.

Section	Description
Pipe characteristics	Information about the physical pipe properties (Ex: Diameter, depth, length)
Emergency repair	Information about the emergency repair (Ex: Immediate Leakage fixes)
Smoke testing assessment	Information about any smoke observed from pipes (Ex: Medium smoke observed emanating from cleanout)
Defects	Information about the pipes using CCTV Cameras (Ex: Multiple Defects)
Composite assessment	Information about the composite material around the pipe
Criticality assessment	Information about the risk value of the pipe (Ex: Medium)
Capacity	Information about the pipe capacity
Summary	Information about total major and minor defects

TABLE 2 PACP incorporated comprehensive rating protocol.

Defect rating	Description
Defect rating 1	Reassess in 10 years.
Defect rating 2	Rehabilitate or replace in 6–10 years.
Defect rating 3	Rehabilitate or replace in 3–5 years.
Defect rating 4	Rehabilitate or replace in 0–2 years.
Defect rating 5	Rehabilitate or replace immediately.

2. Methods and materials

2.1. Data set and data preprocessing

A total of 3100 pipe repair documents were extracted from the Department of Engineering and Environmental Services' approved database by removing records with insufficient and missing information for further analysis. There was no complete information about the defect location, so 130 documents were removed and finally had 2,970 pipe repair documents. Table 1 shows the information included in the pipe repair document.

A Pipeline Assessment Certification Program (PACP) version 7.0.4, released on October 1, 2020, incorporated a Comprehensive rating protocol was established to provide a standardized way of documenting features and assigning defect ratings during the inspection to schedule maintenance. In addition to the PACP defect ratings, numerous other factors such as sewer pipe diameter, pipe material, burial depth, pipe bedding, load transfer, pipe joint type and material, surface loading, ground conditions, groundwater level, and soil type, type of waste carried, pipe age, sediment level, surcharge, and poor maintenance practices were assessed to provide a more precise assessment and these Rating, and it is listed as comprehensive rating protocol which is followed in the maintenance records (Vladeanu and Matthews, 2019a; Betgeri et al., 2023a). Comprehensive and PACP allows wastewater professionals to classify, evaluate and manage inspection data. The most used defect scale is the Comprehensive and PACP Protocol manual in Table 2.

Next, Data preprocessing for the text from pipe repair documents is performed. Data preprocessing is crucial when dealing with text data because text data is unstructured data and the interpretations of the documents by different inspectors. Next, all special characters in the pipe repair documents (e.g., commas in a list) are removed. Misprint in the dataset included typographical errors (e.g., "Leaks" instead of "Laeks"), and NLTK handles the correction of the spelling errors. Then, boundary detection using Stanford parser's sentence detection is performed since it is a reasonably accurate tool in NLP (Malek Mohammadi et al., 2019). Pipe repair documents contain many negation statements (no leaks, no defects... etc.). To identify such terms, entities, or sentences, the Negex algorithm, which is a Python module, is used for negation term detection (Toutanova et al., 2003).

2.2. Annotation

Next, Standard documents are manually annotated for entities and pipe defect ratings. Manual annotation of pipe repair documents provided a gold standard to benchmark the proposed model for pipe defect rating. Two annotators manually mark the list of lexical units and assign a defect rating using the Comprehensive and PACP Protocol to each record. The meaning of sentences and defect ratings can be interpreted differently by different ways inspectors. Annotating final size 2970 pipe repair documents at the entity and record levels and calculating the kappa coefficient for each document and entity. If there is any disagreement between the experts, documents must be annotated again, which is a time taking process. So, a stratified random selection of 500 records is considered a golden standard, annotating each document at the entity and document level. A statistical measure of Cohen's kappa coefficient is used to find the agreement between the two-annotator in Inter Annotator Agreement (IAA) (Kiliç, 2015). Inter Annotator Agreement for our entities is shown in Table 3. Kappa coefficients are considered for evaluating agreement coefficients between the expert's opinions compared for each entity level and record level annotation. So, a stratified random selection of 500 pipe repair documents was selected for the proposed model. Most of our pipe repair documents have defect ratings of 3 and 4. To have an equal number of defect ratings for train and testing our model, 500 pipe repair documents are only selected. 80% of the pipe repair documents were used as training data and 20 percent as test data to evaluate the proposed framework for defect rating, as shown in Figure 1.

2.3. Lexicon generation

Knowledge bases for pipe defects are built by collecting information from documents and structured sources. The creation of a general list of defect attributes is referred to as a defect lexicon or seed words. Secondly, a list of defects was hand-picked for inspectors from traditional sources. After manually selecting defect

TABLE 3 Annotator agreement table.

Entities and defect rating	IAA
Defect	0.78
Size defects	0.82
Locations of defect	0.84
Defect rating 1	0.75
Defect rating 2	0.79
Defect rating 3	0.84
Defect rating 4	0.83
Defect rating 5	0.92

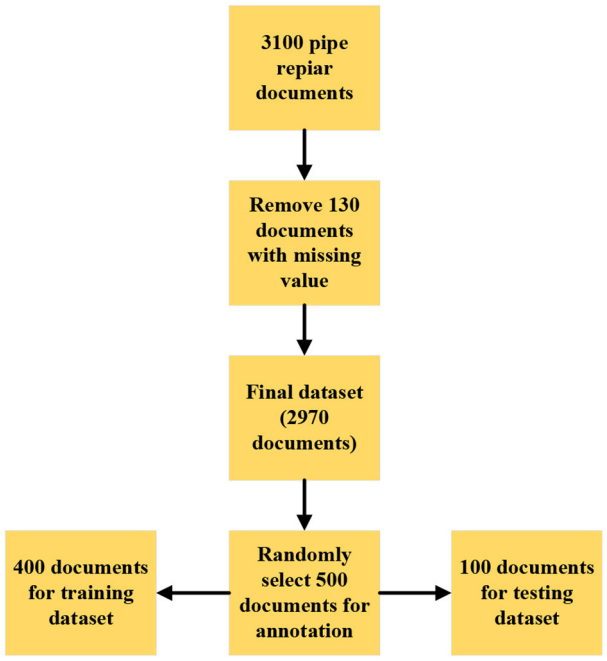

FIGURE 1
Training and testing data selection.

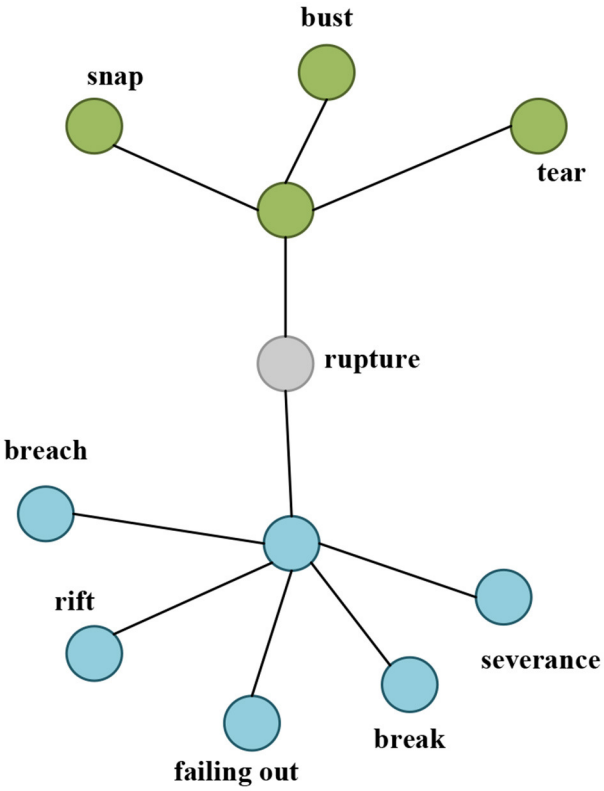

FIGURE 2
Sample knowledge base construction for defect term "Rupture."

attributes, each attribute was expanded with the appropriate verb, noun, and adjective where possible, e.g., leak, leaking, leaked. Then, a knowledge-based method for identifying the new defect attributes is used. Knowledge-based methods exploit available lexicographical resources such as WordNet or HowNet. A lexicon was developed by searching WordNet for a term's synonyms and antonyms (Hu and Liu, 2004). According to Kamps et al. (2004), the closer two words means fewer iterations are needed to identify the synonymous connection between those two words. The relationship between terms in a knowledge base was employed in both investigations. As illustrated in Figure 2, these systems' basic technique is to use seed sets of pipe fault terms and their orientations to expand this collection of defect characteristics by searching for synonyms and antonyms in a knowledge base.

There were a lot of synonyms that needed to be more on-topic and unconnected. After manual tracking of individual seed keywords, a problem is discovered. Their synonyms revealed ambiguous synonyms such as homonyms and antagonyms and incorrect or fuzzily interpreted synonyms. By making different blacklists for each synonym, including redundant, misunderstood, and fuzzily understood terms, the problem is fixed. Unnecessary processing of undesirable terms is avoided in this way. The lexicon knowledge file of pipe defect ontology is shown in Table 4.

2.4. Entity extraction

Entity Extraction aims to identify entities mentioned in the text and classify them into predefined entity types, as shown in Table 5. Manual rules are created to fix the problem of dealing with unstructured pipe repair documents from multiple inspectors. Sentences and specific items from the text employed in defect rating calculations are to be identified; for example, "Leak" should be recognized as a defect. Entity extraction graphical representation is shown in Figure 3.

Lastly, a Bi-LSTM neural network is implemented. The output vector from both forward and backward sequences is adjoined to obtain the final entity representation vector using the lexicon file generated in section 3.1. The Bi-LSTM model is composed of two LSTM networks and is capable of reading input reviews

in both directions, forward and backward. The forward LSTM processes information from left to right and its hidden state and it is shown $\overrightarrow{h}_t = LSTM\ (x_t, \overrightarrow{h}_{t-1})$ and the backward LSTM processes information by reading from right to left and its hidden state can be expressed as $\overleftarrow{h}_t = LSTM\ (x_t, \overleftarrow{h}_{t-1})$. Finally, the output of Bi-LSTM can be summarized by concatenating the forward and backward states as $h_t = [\overrightarrow{h}_t, \overleftarrow{h}_t]$ and the data frame for each sentence consisting of information such as (defects, size of defect, location of defect, and frequency of defects) are created, which will be used for defect rating calculation. The structure of a Bi-LSTM network is shown in Figure 4. The parameter settings for the entity extraction model are shown in Table 6.

TABLE 4 Description of pipe defect ontology.

Location	Frequency of defects	Defects
Mid-point, Upstream, Downstream, Depth category, Pipe length at longitudinal, Spiral, Circumferential,	Rarely, Several, Frequently, Often, Moderate, Very rarely	Fractures, Sags, Smoke, Leaks, Water level, Sags, Deposits, Joint offset, Deposits attached Encrustation, Deposits settled, compacted, Infill runner, Intruding sealing hanging, Intruding sealing ring loss/poorly fitting, Tap factory defective, Corrosion, Pitting, Gap, Hole, Stain, Rough spot, Foible, Rupture, No defect

2.5. Defect rating

The proposed defect rating calculation employs three aspects of within pipe repair document defect term frequency ($w_{frequencies}$), the importance of a defect term ($w_{defects}$), and the location of the defect ($w_{location}$) within the pipe, which is developed using term frequency algorithm (Azam, 2012).

The location of the defect plays an important role in defect rating, which mentions if the defect mentioned is in one particular location or multiple locations. For locations, the weights are assigned based on the scores mentioned in Pipeline Assessment - NASSCO (Cosham and Hopkins, 2004; Lepot et al., 2017; Wang, 2021). Table 7 shows the randomly assigned weights based on the pipe defect location. When no location is found $w_{location} =$

TABLE 5 Description of the entities.

Entities	Description
Defect	Keywords (e.g., Leakage, Rupture, etc.)
Location of defect	Keywords (e.g., junction, end, etc.) or distance from the end of the pipe
Frequency of defects	Keywords (e.g, Rarely, Frequently, etc.)

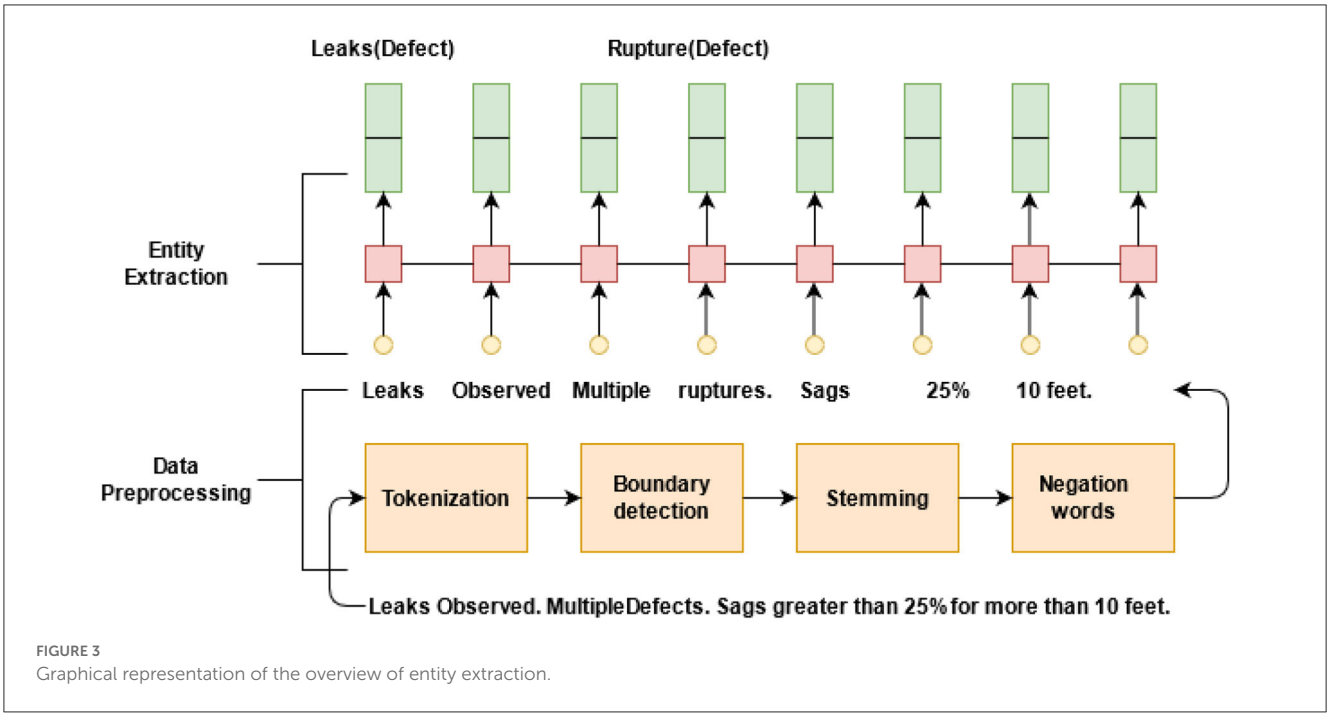

FIGURE 3
Graphical representation of the overview of entity extraction.

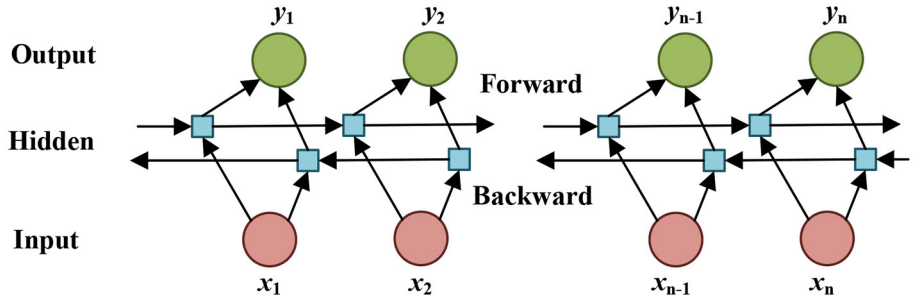

FIGURE 4
Layers in a Bi-LSTM neural network.

TABLE 6 Parameter setting of the entity extraction model.

Parameter	Value
Word vector embedding size	200
Dictionary feature vector embedding size	100
# Hidden neurons for each hidden layer	300
Batch size	100
Tag Indices	4
Learning Rate	0.005
Number of epochs	10
Optimizer	Adam optimizer

TABLE 7 Weights assignment based on the location of the defect.

Location	Assigned weight
One location	$w_{location} = 0.9$
Multiple locations	$w_{location} = 1.0$
No location	$w_{location} = 1.0$

TABLE 8 Weights assignment based on the frequency of the defects outcome.

Frequency outcome	Assigned weight
Very rarely or none	$w_0 = 0.1$
Rarely	$w_1 = 0.25$
Moderate	$w_2 = 0.50$
Moderate to Frequently	$w_3 = 0.75$
Frequently, More or very Frequently/ Several/ Oftenly	$w_4 = 0.99$

1.0 random weight is assigned because the exact location of the defect was not mentioned. Table 8 shows the randomly assigned weights based on the defect occurrence frequency. Similarly, the defects factor is randomly chosen based on the pipe failure. Table 9 shows the random weights assignment based on defect lexicon units found. Table 10 shows the defect ratings assigned based on $w_{frequencies}$, $w_{location}$, w_{defect}.

TABLE 9 Weights assignment based on lexicon units found.

Lexicon or defect found	Assigned weight
No lexicon unit	$w_{defect} = 0.5$
One lexicon unit	$w_{defect} = 0.8$
Multiple lexicon unit	$w_{defect} = 1.0$

TABLE 10 Defect rating assignment based on $w_{frequencies}$, $w_{location}$, w_{defect}.

Defect rating	$w_{frequencies}$, $w_{location}$, w_{defect}
Defect rating 1	$w_{frequencies} = 0.1$, $w_{location} = 0.9$ or 1.0, $w_{defect} = 0.5$
Defect rating 2	$w_{frequencies} = 0.25$, $w_{location} = 0.9$ or 1.0, $w_{defect} = 0.8$ or 1.0
Defect rating 3	$w_{frequencies} = 0.5$, $w_{location} = 0.9$ or 1.0, $w_{defect} = 0.8$ or 1.0
Defect rating 4	$w_{frequencies} = 0.75$, $w_{location} = 0.9$ or 1.0, $w_{defect} = 0.8$ or 1.0
Defect rating 5	$w_{frequencies} = 0.99$, $w_{location} = 0.9$ or 1.0, $w_{defect} = 0.8$ or 1.0

Defect Rating Score example, 05CCD pipe inspection document contains the following information, as shown in Box 1.

It consists of only one sentence, and the location factor is 0.9 because it has only one location. The $w_{frequencies}$ term weights 0.99 for w_4, which matches very frequently. The defect term (leakage) matches with the lexicon unit and has a 1.0 (seed term) weight, so the w_{defect} is 0.8

So, for the example shown, the rating assigned for Pipe 05CCD is 5, which means it needs an immediate replacement or rehabilitation.

Figure 5 illustrates the entire schema of the proposed framework.

3. Results and discussions

The proposed framework defect rating model was trained on the training subset. Finally, the trained models were evaluated on the test subset. To test the defect rating model, precision, recall,

BOX 1 Sample pipe repair document.

Very Frequently, there is a leakage in pipe 10 feet away from the pipe installation

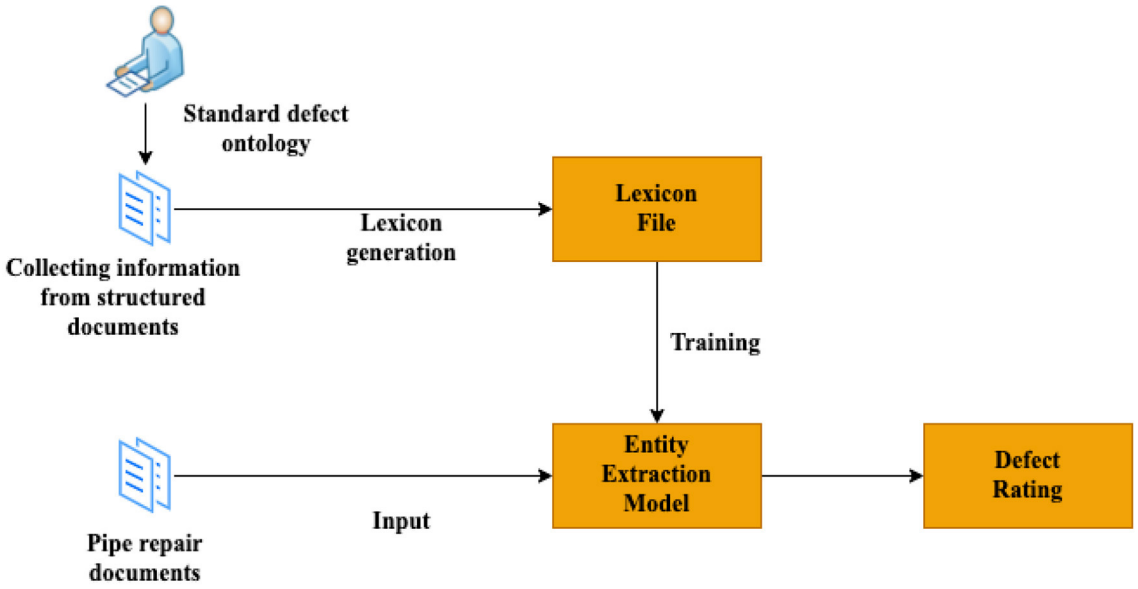

FIGURE 5
A framework for the proposed defect rating.

accuracy, specificity, and F1 score presented in Equations 1–5 were utilized (Sokolova et al., 2006).

$$\text{Accuracy} = \frac{\text{True positive} + \text{True Negative}}{\text{True positive} + \text{True Negative} + \text{False Positive} + \text{False Negative}} \quad (1)$$

$$\text{Precision} = \frac{\text{True positive} + \text{True Negative}}{\text{True positive} + \text{False Positive}} \quad (2)$$

$$\text{Recall} = \frac{\text{True positive}}{\text{True positive} + \text{False Negative}} \quad (3)$$

$$\text{Specificity} = \frac{\text{True Negative}}{\text{True positive} + \text{False Negative}} \quad (4)$$

$$\text{F1 Score} = 2 \times \frac{\text{Recall} \times \text{Precision}}{\text{Precision} + \text{Recall}} \quad (5)$$

To accurately assess the quality of pipes for scheduling maintenance, it is essential to extract detailed information from pipe repair documents using entity extraction. The entity extraction model identifies pipe repair attribute entities from the pipe repair documents. The entity extraction model's overall average accuracy for all entities is 95.1%. The F1 scores of defects, defect size, location of the defect, and frequency of defect terms/entities are 93.6%, 92.4%, 93.6%, and 95.2%, as shown in Table 11. The entity extraction calculates the defect scores and assigns the defect rating.

The defect statements' representation varied from inspection to inspection. Pipes, for example, can be represented in a variety of ways, such as a "pipeline," a "sewer Line," or a "waterline." The entity extraction model adapted well to different inspector reporting styles and languages. Findings show that accuracy, precision, F1-score, and recall have all improved significantly. The entity extraction accuracy of the Bi-LSTM entity model was consistently higher. Experiments revealed that using sentence boundary detection, entities, and a domain dictionary in the tasks improved accuracy considerably, demonstrating the use of merging techniques and domain dictionaries in entity extraction tasks.

A defect rating must be correct for it to be considered accurate. Precision, recall, accuracy, and F1 score were calculated to evaluate the models' performance. The proposed defect rating model achieved higher accuracy by 97.0%, 92.0%, 93.0%, 95.0%, and 98.0% for the defect rating 1, defect rating 2, defect rating 3, defect rating 4, and defect rating 5, respectively, as shown in Table 9. After a deep analysis of the results, a conclusion was made that misclassified records have complicated sentences, which are very hard for the model to understand and classify, For example. multiple is used for both defects and locations. Table 12 shows the results of the defect rating assignment.

The results indicate that defect rating can be accurately calculated with the help of the entity extraction model. However, the defect rating model struggled to accurately calculate defect rating pipe repair documents due to fewer defect attributes in pipe repair documents, high disagreement between inspectors who annotated pipe repair documents, and not mentioning the frequency of the defect in a few pipe repair documents.

TABLE 11 Results of entity extraction model units (%).

Entity tags	Accuracy	Recall	Specificity	Precision	F1
Defects	92.0	94.1	92.3	95.0	93.6
Location of defect	95.0	91.1	96.9	97.5	93.6
Frequency of defects	96.2	95.6	93.0	94.2	95.2

TABLE 12 Results for defect rating model units (%).

Defect rating score	Accuracy	Recall	Specificity	Precision	F1
Defect rating 1	97.0	97.5	97.0	97.0	97.5
Defect rating 2	92.0	93.0	92.5	94.0	93.0
Defect rating 3	93.0	94.0	95.5	95.5	95.5
Defect rating 4	95.0	93.0	92.5	95.0	94.5
Defect rating 5	98.0	97.0	97.5	98.0	98.0

Waste water pipe maintenance is a critical issue faced by many utilities in the United States. PACP protocol, PACP incorporated Comprehensive rating protocol was developed a few years back to resolve the issue by the utilities. Both developed methods follow manual inspection. The defect occurrence frequency is a vital determinant in determining the severity of the pipe failure. The variety of defects essentially invalidates the assumptions of text mining approaches such as decision trees because a single defect could be divided into multiple categories.

Frequency as an indicator of severity helps to favor the automation process such as tf-idf (Robertson, 2004; Zhang et al., 2011). tf-idf determines the uniqueness of a text through frequently or rarely. tf-idf would detect the pipe failure but not the severity of the pipe failure and tf-idf. In our approach, the words often, frequently, rarely, etc capture the frequency of the defect. Negation and Location play a vital role not considering them could create data noise.

A disadvantage of the Machine Learning method approach is focusing on correlations, which means results are based on statistics. An inspector may have difficulty understanding the correlations; however, a grammatical approach used in this work can be more understandable.

Compared to Multi-Criteria Decision Analysis (MCDA) methods, our proposed approach has multiple advantages. MCDA method is very sensitive to data; interdependence between criteria and alternatives can lead to inconsistent judgments, whereas NLP deals with 6000 languages and can identify any data. The MCDA method requires employing human staff, which is costly. In contrast, our proposed model is less costly because it employs no human staff.

4. Strength and limitations

This study is groundbreaking in several ways. This is the first study to look at the validity of NLP for numerous defect entities while investigating pipe repair papers for defecting ratings. Secondly, NLP research in pipe repair documents concentrates exclusively on a single defect condition, such as leaks. With the help of lexicon generation using WordNet, a completed set of defect Lexicon is created and used in defect rating. There are multiple drawbacks also to the proposed model. Firstly, To address accurate defect detection and recognition of wastewater defects that deal with pipe assessment document data, an approach for detecting and recognizing defects, including pipe geographical location, is essential and is not considered in the proposed model. Secondly, the present level of wastewater system management differs from city to city. The algorithm must be trained to improve by collecting more pipe-related lexicon terms from different cities and must be applied to different cities' data. Thirdly, the dataset had less defect ratings related to 1, 2, and 5, so to improve the model, a larger data set with an equal number of defect ratings are needed.

5. Conclusions and future work

Since the health state of the wastewater pipe is assessed and recorded as the basis for decision-making in this process, wastewater pipe assessment is the foundation for creating an effective maintenance plan. The proposed model is a reliable and accurate approach to detecting pipe repair documents and assigning defect ratings. The proposed model is suitable for actual pipe repair documents —in terms of performance, inference speed, and simplicity in output interpretation, which can accurately characterize a particular location and defect. This model is tailored to solve the challenges in the wastewater infrastructure. The lexicon generation step is the core for data integration to the proposed model, where the defect ontology entities and semantic rules are developed for representing different types of information related to the specific location, which helps the user to customize the defect lexicon according to location. However, more study is needed to assess the applicability and validity of NLP approaches for trenchless or building projects. In the future, there is a chance to improve the proposed algorithm by collecting more pipe-related lexicon terms from data from different cities and adding pipe-specific geographical locations. Secondly, a framework that consists of data extractions of CCTV videos using Deep Learning Algorithms and feeds the data for assessing the Suitability of

Trenchless Technologies by the Decision Support System for installing, replacing, or rehabilitating each pipe and reducing the methods of costs and evaluating all the 70 technologies and helps in selecting the appropriate techniques would be developed.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

SB led this paper and others contributed to the outcomes. All authors contributed to the article and approved the submitted version.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Azam, N. (2012). and J. Yao, Comparison of term frequency and document frequency based feature selection metrics in text categorization. *Exp. Syst. Appl.* 39, 4760–4768. doi: 10.1016/j.eswa.2011.09.160

Betgeri, S. N. (2022). *Analytic Hierarchy Process is not a Suitable method for the Comprehensive Rating (Doctoral* dissertation, Louisiana Tech University).

Betgeri, S. N., Matthews, J. C., and Vladeanu, G. (2023a). Development of comprehensive rating for the evaluation of sewer pipelines. *J. Pipeline Sys. Eng. Practice* 14, 04023001. doi: 10.1061/JPSEA2.PSENG-1208

Betgeri, S. N., and Smith, D. B. (2021). *Comparison of Sewer Conditions Ratings with Repair Recommendation Reports.* North American Society for Trenchless Technology (NASTT). Available online at: https://member.nastt.org/products/product/2021-TM1-T6-01

Betgeri, S. N., Vadyala, S. R., Matthews, J. C., Madadi, M., and Vladeanu, G. (2023b). Wastewater pipe condition rating model using K-nearest neighbors. *Tunnell. Underg. Space Technol.* 132, 104921. doi: 10.1016/j.tust.2022.104921

Boskabadi, A., Mirmozaffari, M., Yazdani, R., and Farahani, A. (2022). Design of a distribution network in a multi-product, multi-period green supply chain system under demand uncertainty. *Sust. Oper. Comput.* 3, 226–237. doi: 10.1016/j.susoc.2022.01.005

Caldas, C. H., and Soibelman, L. (2003). Automating hierarchical document classification for construction management information systems. *Autom. Constr.* 12, 395–406. doi: 10.1016/S0926-5805(03)00004-9

Cambria, E., and White, B. (2014). Jumping NLP curves: A review of natural language processing research. *IEEE Comput. Int. Mag.* 9, 48–57. doi: 10.1109/MCI.2014.2307227

Chahinian, N., Bonnabaud La Bruyère, T., Frontini, F., Delenne, C., Julien, M., Panckhurst, R., et al. (2021). WEIR-P: An information extraction pipeline for the wastewater domain. In *International Conference on Research Challenges in Information Science*. Cham: Springer International Publishing, 171–188

Cheng, J. C., and Wang, M. (2018). Automated detection of sewer pipe defects in closed-circuit television images using deep learning techniques. *Autom. Constr.* 95, 155–171. doi: 10.1016/j.autcon.2018.08.006

Chi, N. W., Lin, K. Y., and Hsieh, S. H. (2014). Using ontology-based text classification to assist Job Hazard Analysis. *Adv. Eng. Inf.* 28, 381–394. doi: 10.1016/j.aei.2014.05.001

Cosham, A., and Hopkins, P. (2004). An overview of the pipeline defect assessment manual (PDAM). *Int. Pipeline Technol. Conf.* 29, 720–745.

Dang, L. M., Hassan, S. I., Im, S., Mehmood, I., and Moon, H. (2018). Utilizing text recognition for the defects extraction in sewers CCTV inspection videos. *Comput. Ind.* 99, 96–109. doi: 10.1016/j.compind.2018.03.020

Endalie, D., Haile, G., and Taye, W. (2022). Bi-directional long short term memory-gated recurrent unit model for Amharic next word prediction. *PloS ONE* 17, e0273156. doi: 10.1371/journal.pone.0273156

Graves, A., Mohamed, A. R., and Hinton, G. (2013). "Speech recognition with deep recurrent neural networks," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, Manhattan, NY: IEEE, 664—6649.

Hassan, S. I., Dang, L. M., Mehmood, I., Im, S., Choi, C., Kang, J., et al. (2019). Underground sewer pipe condition assessment based on convolutional neural networks. *Autom. Constr.* 106, 102849. doi: 10.1016/j.autcon.2019.102849

Hu, M., and Liu, B. (2004). Mining and summarizing customer reviews. *ACM SIGKDD* 12, 168–177. doi: 10.1145/1014052.1014073

Jallan, Y. (2020). *Text Mining of the Securities and Exchange Commission Financial Filings of Publicly Traded Construction Firms Using Deep Learning to Identify and Assess Risk.* Georgia: Institute of Technology. doi: 10.1061/(ASCE)CO.1943-7862.0001932

Kamps, J., Marx, M., Mokken, R. J., and Rijke, D. M. (2004). Using WordNet to measure semantic orientations of adjectives. *LREC* 4, 1115–1118.

Kiliç, S. (2015). Kappa testi. *J. Mood Disorders* 5, 142–144. doi: 10.5455/jmood.20150920115439

Le, T., and David Jeong, H. (2017). NLP-based approach to semantic classification of heterogeneous transportation asset data terminology. *J. Comput. Civil Eng.* 31, 04017057. doi: 10.1061/(ASCE)CP.1943-5487.0000701

Lepot, M., Stani,ć, N., and Clemens, F. H. (2017). A technology for sewer pipe inspection (Part 2): Experimental assessment of a new laser profiler for sewer defect detection and quantification. *Autom. Constr.* 73, 1–11. doi: 10.1016/j.autcon.2016.10.010

Li, D., Cong, A., and Guo, S. (2019). Sewer damage detection from imbalanced CCTV inspection data using deep convolutional neural networks with hierarchical classification. *Autom. Constr.* 101, 199–208. doi: 10.1016/j.autcon.2019.01.017

Li, S., Cai, H., and Kamat, V. R. (2016). Integrating natural language processing and spatial reasoning for utility compliance checking. *J. Constr. Eng. Manage.* 142, 04016074. doi: 10.1061/(ASCE)CO.1943-7862.0001199

Malek Mohammadi, M., Najafi, M., Kaushal, V., Serajiantehrani, R., Salehabadi, N., Ashoori, T., et al. (2019). Sewer pipes condition prediction models: a state-of-the-art review. *Infrastructures* 4, 64. doi: 10.3390/infrastructures4040064

Marcus, M. (1995). New trends in natural language processing: statistical natural language processing. *Proc. Nat. Acad. Sci.* 92, 10052–10059. doi: 10.1073/pnas.92.22.10052

Mohammadi, M. M., Najafi, M., Tabesh, A., Riley, J., and Gruber, J. (2019). *Condition Prediction of Sanitary Sewer Pipes. Pipelines*. Reston, VA: American Society of Civil Engineers, 117–126.

Moradi, S., Zayed, T., Nasiri, F., and Golkhoo, F. (2020). Automated anomaly detection and localization in sewer inspection videos using proportional data modeling and deep learning–based text recognition. *J. Inf. Syst.* 26, 04020018. doi: 10.1061/(ASCE)IS.1943-555X.0000553

Nicklow, J., Reed, P., Savic, D., Dessalegne, T., Harrell, L., Chan-Hilton, A., et al. (2010). State of the art for genetic algorithms and beyond in water resources planning and management. *J. Water Res. Plan. Manage.* 136, 412–432. doi: 10.1061/(ASCE)WR.1943-5452.0000053

Niu, S., and Srivastava, V. (2022). Ultrasound classification of interacting flaws using finite element simulations and convolutional neural network. *Eng. Comput.* 38, 4653–4662. doi: 10.1007/s00366-022-01681-y

Robertson, S. (2004). Understanding inverse document frequency: on theoretical arguments for IDF. *J. Docum.* 60, 503–520. doi: 10.1108/00220410410560582

Shafiei Alavijeh, M., Scott, R., Seviaryn, F., and Maev, R. G. (2021). Using machine learning to automate ultrasound-based classification of butt-fused joints in medium-density polyethylene gas pipes. *The J. Acoust. Soc. Am.* 150, 561–572. doi: 10.1121/10.0005656

Sokolova, M., Japkowicz, N., and Szpakowicz, S. (2006). "Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation," in *Australasian Joint Conference on Artificial Intelligence,* Berlin, Heidelberg: Springer Berlin Heidelberg, 1015–1021

Toutanova, K., Klein, D., Manning, C. D., and Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. *Pro. Assoc. Comput. Ling.* 18, 252–259. doi: 10.3115/1073445.1073478

Tscheikner-Gratl, F., Caradot, N., Cherqui, F., Leitão, J. P., Ahmadi, M., Langeveld, J. G., et al. (2019). Sewer asset management–state of the art and research needs. *Urban Water J.* 16, 662–675. doi: 10.1080/1573062X.2020.1713382

Vadyala, S. R., and Betgeri, S. N. (2021). *Predicting the spread of COVID-19 in Delhi, India using Deep Residual Recurrent Neural Networks*. arXiv preprint arXiv:2110.05477

Vadyala, S. R., Betgeri, S. N., and Betgeri, N. P. (2022a). Physics-informed neural network method for solving one-dimensional advection equation using PyTorch. *Array* 13, 100110. doi: 10.1016/j.array.2021.100110

Vadyala, S. R., Betgeri, S. N., Matthews, J. C., and Matthews, E. (2022b). A review of physics-based machine learning in civil engineering. *Res. Eng.* 13, 100316. doi: 10.1016/j.rineng.2021.100316

Vadyala, S. R., Betgeri, S. N., Sherer, E. A., and Amritphale, A. (2021). Prediction of the number of COVID-19 confirmed cases based on K-means-LSTM. *Array* 11, 100085. doi: 10.1016/j.array.2021.100085

Vadyala, S. R., and Sherer, E. A. (2021). *Natural Language Processing Accurately Categorizes Indications, Findings and Pathology Reports From Multicenter Colonoscopy*.

Vladeanu, G., and Matthews, J. (2019b). Wastewater pipe condition rating model using multicriteria decision analysis. *J. Water Res. Plan. Manage.* 145, 04019058. doi: 10.1061/(ASCE)WR.1943-5452.0001134

Vladeanu, G. J., and Matthews, J. C. (2019a). Consequence-of-failure model for risk-based asset management of wastewater pipes using AHP. *J. Pip. Syst. Eng. Prac.* 10, 04019005. doi: 10.1061/(ASCE)PS.1949-1204.0000370

Wang, M. (2021). Ontology-based modelling of lifecycle underground utility information to support operation and maintenance. *Automat. Constr.* 132, 103933. doi: 10.1016/j.autcon.2021.103933

Yang, L., and Zhao, Q. (2020). A novel PPA method for fluid pipeline leak detection based on OPELM and bidirectional LSTM. *IEEE Access* 8, 107185–107199. doi: 10.1109/ACCESS.2020.3000960

Yildirim, Ö. (2018). A novel wavelet sequence based on deep bidirectional LSTM network model for ECG signal classification. *Comput. Biol. Med.* 96, 189–202. doi: 10.1016/j.compbiomed.2018.03.016

Yugandhar, V., and Nethra, B. S. (2014). *Statistical Software Packages for Research in Social Sciences. Recent Research Advancements in Information Technology*. Berlin: Springer.

Zhang, W., Yoshida, T., and Tang, X. (2011). A comparative study of TF* IDF, LSI and multi-words for text classification. *Exp. Syst. Appl.* 38, 2758–2765. doi: 10.1016/j.eswa.2010.08.066

Zhong, B., Xing, X., Love, P., Wang, X., and Luo, H. (2019). Convolutional neural network: Deep learning-based classification of building quality problems. *Adv. Eng. Inf.* 40, 46–57. doi: 10.1016/j.aei.2019.02.009