YAKE-Guided LDA approach for automatic classification of construction safety reports

H. Gadekar[a] and N. Bugalia[a]

[a] Department of Civil Engineering, Indian Institute of Technology Madras, India
E-mail: hrishikesh.gadekar89@gmail.com, nikhilbugalia@gmail.com

Abstract –

Identifying efficient processes for classifying text-based safety reports using Machine-Learning (ML) is an essential area of research. However, much of the previous work on the topic relies on supervised learning approaches, which are often manually intensive and require large volumes of pre-labeled data. To achieve reduced requirements for human intervention during the classification process, the current study tests the applicability and validity of a state-of-the-art unsupervised learning approach, i.e., Yet Another Keyword Extractor (YAKE) integrated with Guided Latent Dirichlet allocation (GLDA). The current study is the first known application of the approach for the construction sector. Web-based, readily accessible information is used to develop a domain corpus. The keywords obtained from the domain corpus using YAKE are seeded in GLDA to classify nearly 13,000 safety reports from two different datasets in 4 commonly used category labels. The study demonstrates that moderate to high classification performance is achievable through the YAKE-GLDA approach. A high F1 score of 0.82 for the Personal-protective equipment category and a total F1 score of 0.62 is achievable. Furthermore, the same domain corpus helps achieve good classification performance across different datasets, highlighting the generalizability of the YAKE-GLDA approach. However, results from novel sensitivity analysis show a non-generalizable trend for sensitivity to hyperparameters. Hence attention is warranted for potential consistency issues facing the approach. The preliminary results demonstrate outstanding potential for the YAKE-GLDA approach for wide-ranging adoption in the construction industry. However, future work should also focus on more granular classification labels applications and improving classification efficiency.

Keywords –

Unsupervised machine learning; Construction safety; Text classification; Topic modeling

1 Introduction

Despite a significant improvement over the years, the construction sector, compared to other sectors, continues to perform poorly for issues relating to Occupational Health and Safety (OHS) [1]. To improve safety performance, one of the central ideas is to collect large volumes of safety observations (SOs), such as accidents, injuries, and near-miss reports, and utilize these reports to enable organizational learning [2].

However, literature has also highlighted construction organizations' challenges in sustaining safety reporting and organizational learning efforts, often due to their resource intensiveness [2]. For example, the safety observations in the construction sector are often unstructured textual data and of poor quality requiring extensive manual efforts to process such information [2]. Therefore, for solving practical issues faced by the construction organizations, identifying efficient processes for the classification of the text-based information (such as the SOs) using Machine-Learning (ML) and text-mining-based approaches continues to be an essential area of research [3,4].

Across domains, including construction, a significant proportion of literature focussing on ML-based classification of SOs continues to rely on supervised ML approaches [4]. Many studies have demonstrated the high classification efficiency of such supervised approaches [3,4]. However, lack of generalizability, the necessity of a large quantity of pre-labeled data, and significant manual inputs during ML-based analysis continue to be a limitation for the broader application of such approaches in actual practice [3,4]. On the other hand, literature exploring unsupervised and semi-supervised approaches is relatively scarce [4,5]. In principle, such approaches can reduce the requirement of human intervention [6].

The key motivation for this paper is to achieve the reduced requirement of human intervention during the classification process while also achieving good classification performance. Consequently, the objective of the current study is to test the applicability and validity of a recently developed unsupervised learning approach,

i.e., Yet Another Keyword Extractor (YAKE) integrated with Guided Latent Dirichlet allocation (GLDA), for classifying construction SOs. The YAKE-GLDA approach is recent and claimed to be a domain-independent approach that has been shown to achieve good classification accuracy with reduced manual efforts [7]. To the best of the authors' knowledge, the current study is the first-ever attempt at the YAKE-GLDA approach for the construction sector [5]. The study makes essential contributions to evaluating the potential of an ML approach in automating SO labeling for the construction sector from an unstructured corpus with minimal manual intervention. Unlike previous studies, the current study also presents the results from sensitivity analysis and hence contributes to state-of-the-art literature on YAKE-GLDA.

The study is structured as follows. Section 2 provides an overview of the literature and identifies the essential gaps where the study contributes. Section 3 describes the integrated YAKE-GLDA method and the analytical methodology adopted in the current study to classify the SOs in construction from two different data sources. Results have been summarized in section 4, followed by discussions in section 5. Conclusions have been summarized in section 6.

2 Literature Review

For the construction sector, literature focusing on analyzing textual SOs using ML techniques using unsupervised learning approaches has been relatively scarce but has been growing [4]. Based on a state-of-the-art literature review on the topic, three broad unsupervised learning approaches are essentially used. These are (a) *Associated Rule Mining* (ARM) approaches, (b) *Text-mining* approaches, and (c) *Clustering* techniques [4].

The primary purpose of ARM approaches is to find the associations or relationships (called rules) among the input variables toward the defined outcomes [4]. ARM's functionality has also been extended toward text classification [8]. However, many challenges facing ARM approaches are that typically large quantities of rules containing many parameters get generated, which are difficult to comprehend for human interpretations [9]. Due to such issues, their applications in classification remain limited. On the other hand, the unsupervised *Text-Mining* approaches have also been used as pre-processing steps [4,10]. They are highly efficient in converting the typically unstructured data available in construction SOs to structured data leading to very high classification performance [10]. However, the overall process is highly resource-intensive, requiring intensive manual intervention and domain expertise to formulate and validate rules [3]. There is also a possibility that the rules thus developed are specific to one type of data source and not generalizable to other data sources, even within the construction sector. Hence, ARM and Text-mining approaches are not ideally suited for reducing the manual intervention.

On the other hand, *Clustering* is a frequently used unsupervised learning approach useful in grouping the data into different clusters or topics having a substantial similarity among the members belonging to each topic [5]. LDA and K-means clustering are frequently used clustering approaches [4,5]. Clustering approaches are often fully automated, thereby reducing the necessity of manual inputs to a great extent. Studies focusing on construction SOs have also shown the value-addition of various clustering approaches in obtaining meaningful topics relevant to safety [6].

However, there are significant limitations with conventional clustering approaches for classification tasks. Since the clustering process is fully automated, the optimal number of clusters may not always match the user requirements of classification labeling. Furthermore, in many cases, the outputs of clustering approaches are not comprehensible to human decision-makers [7]. The clustering process may not always be guided through the relationship patterns commonly understood in a specific domain, such as construction [7]. Several recent studies have attempted to improve upon the limitations of the traditional clustering process, mainly by seeding keywords while initiating the clustering process. Through such a seeding process, the topics generated by the clustering process are human interpretable, and their number can be controlled [11]. For example, GLDA is a technique that allows users to seed classification categories using domain-specific keywords as an improvement over the conventional LDA approach [12].

However, extracting domain-specific keywords to be used as seeds could still be challenging. The conventional approach of relying on domain experts to identify and assign keywords is expensive, time-consuming, and error-prone. Hence, several recent studies have also explored automated approaches for keyword extraction. These techniques often require statistics-based features such as frequency-of-word, distance-of-word, and structural features [7]. Many unsupervised keywords extracting techniques, such as Term Frequency-Inverse Document Frequency (tf-idf) [13], PageRank method [14], and Rapid Automatic Keyword Extraction (RAKE) [15], are efficient in keyword extraction based on the statistical features described above. The most recent keyword extraction approach is YAKE [16], which has been shown to perform significantly better than tf-idf and RAKE across many standard datasets [16,17]. YAKE also does not require linguistic information and thus can be used for any language [16,17], extending its generalizability.

The first known application of an integrated YAKE and GLDA unsupervised approach demonstrated its potential in classifying SOs for Aviation and Chemical industry [7]. For the construction sector, the most recent studies have only adopted LDA-based approaches using the tf-idf keyword extractor [13]. Most advanced keyword extractors such as YAKE and RAKE are yet to be explored for their applications in the analysis of construction SOs[5]. This is a significant research gap that the current study aims to fulfill. Unlike in the Aviation and the Process industry, safety reporting systems are often not well-established and mature in the construction sector. Hence, the proven validity and applicability of such an approach with reduced manual intervention for the construction sector may significantly impact the industry-wide adoption of ML approaches.

Furthermore, little is known about the sensitivity of classification performance for the YAKE-GLDA framework, as the early-stage studies have focused only on demonstrating the application [7]. Understanding such sensitivity is crucial to exploring any analytical approach to understanding its true potential. The current study also aims to address this gap in the literature.

3 Methodology

The overall analytical process adopted in the current study has been summarized in Figure 1. The process entails three main steps (1) Building a domain corpus for each of the four categories and extraction of domain-specific keywords using YAKE, (2) Pre-processing the input datasets containing SOs, and (3) Classification of the SOs using the GLDA approach. Subsequent sections provide details on each of the three steps.

3.1 Domain-specific keywords using YAKE

To induce domain knowledge to the GLDA classification model, a set of construction domain-specific keywords are automatically extracted using YAKE from the domain corpora of each of the target categories. A total of 4 target categories are identified based on the category labels available in the primary dataset utilized in the study. More information on the dataset has been described later (see Table 1). A domain corpus is a collection of text drawn from sources containing information particular to the domain. Various online literature sources, including journal papers, research articles, and web resources, have been utilized to assemble four separate domain corpora comprising 31,788 words. A subset of the used resources is mentioned in Table 1.

The domain corpus is pre-processed prior to the extraction of keywords. The same pre-processing steps are also applied to the input datasets. The main pre-processing steps performed are (1) Lowercasing, (2) Punctuation and numbers removal, (3) Spelling correction using spell check library in the programming language python, (4) Tokenization and stop words removal, (5) Stemming and lemmatization [3,18].

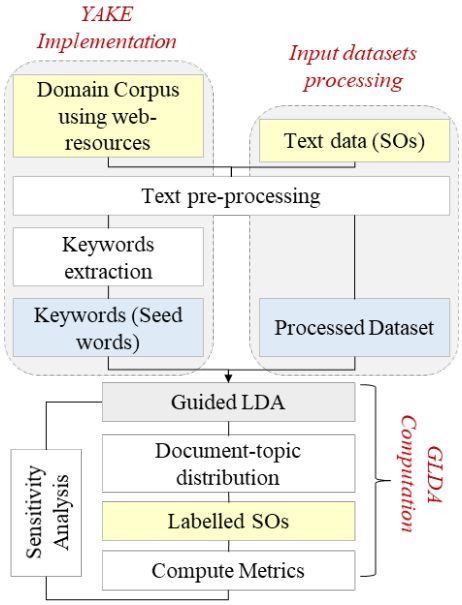

Figure 1. The analytical process of the study

Keywords are a set of content words representing a specific topic. As mentioned earlier, the YAKE method extracts domain-specific keywords from the pre-processed domain corpora. YAKE uses statistical text features extracted from single documents to select relevant keywords. As a document can contain multiple keywords, they are ranked based on their Significance Score SS(k), assigned by the YAKE algorithm. For the computation of SS(k), a set of five features, i.e., Casing (cs), Position (p), Relatedness (r), Frequency (f), and Occurrence (o), is calculated. For brevity, detailed mathematical formulations for these features are not included in the current study, but more information can be found elsewhere [7,16]. The output of YAKE is an ordered list of keywords, ranked based on increasing SS(k), where the smaller the value of SS(k), the more critical the keyword. A set of top 20 keywords for each of the four categories is used for seeding purposes.

3.2 Input datasets and their processing

SOs from two different sources are used in the current study. These datasets represent the diversity of the SOs available in the construction industry.

Dataset 1 is the primary dataset used to demonstrate the validity of the YAKE-GLDA approach for the construction safety domain. *Dataset 1* contains worker-reported near-miss safety observations from a large-scale construction site on a natural gas plant in Kuwait. As per the prevalent reporting practice at this site, the focus is to

promote reporting from the front-end workers as much as possible, rather than obtaining SOs only from safety supervisors [2]. The front-end worker reported data faced several quality-related issues, such as a high proportion of misspelled words, poor sentence structuring, and generally smaller description [18]. Because of such data-quality issues, *Dataset 1* represents the large quantity of usually poor quality SO data generated across the globe. The workers write a brief description of the SO and provide a categorization of the SO into four categories, namely Personal-protective equipment (A), Compliance to safe work (B), Equipment or tools (C), and Housekeeping (D).

Table 1. Overview of Dataset 1

Category labels (% of total data) *	Subcategory label examples	A subset of references used for developing domain corpus
A (21.62)	Ear, eyes, face, hand protection; Harness; respiratory protection	Link 1, Link 2, Link 3
B (49.62)	Electrical, Excavation, Fire safety; working at height; Traffic	Link 5, Link 6, Link 7, Link 8
C (08.14)	Equipment usage and selection; color coding; authorization; tags; maintenance	Link 9, Link 10. Link 11,
D (20.62)	Cleaning; hazardous material management; waste segregation and disposal;	Link 12, Link 13
*Dataset 1 contains a total of 12490 observations		

On the other hand, *Dataset 2* was used to test the generalizability aspects of the YAKE-GLDA approach. The domain keywords obtained for analyzing *Dataset 1* were also used to analyze *Dataset 2*. Another essential category of SOs prominently prevalent in the construction industry, i.e., textual narratives containing descriptions of injuries/fatalities at construction sites often stored in well-managed databases by safety professionals, is included in *Dataset 2*. Compared to the worker-reported data, the data reported by safety professionals is better in quality and contains longer descriptions [18].

This study utilizes a sample of publicly available fatality/injury narratives provided by the USA's Occupational Safety and Health Administration (OSHA). Goh and Ubeynarayana [3] have used 1000 observations from the OSHA database and have labeled them into 11 classification categories. *Dataset 2* used in the current study is a subset of the 1000 labeled observations provided in [3]. To test the generalizability of the YAKE-GLDA approach, harmonization of classification labels between the two datasets has been implemented. Hence, the thirteen label categories from *Dataset 2* were mapped with the four broader category labels available from *Dataset 1* (see Table 2). The mapping was also confirmed by reading the detailed description by the two authors experienced in the construction sector. However, not all 11 labels could be readily mapped with the four labels of *Dataset 1*. The extended description in *Dataset 2* included information on multiple potential causes of the accidents/injuries, which could be mapped to different causes of category labels in *Dataset 1* [3]. Hence, only mappings where a precise one-on-one mapping could be obtained have been included in the analysis to avoid inducing errors in the mapping process. As a result, 823 observations out of 1000 available were included in *Dataset 2*.

Table 2. Label Harmonization between datasets 2 and 1

Labels in *Dataset 2*	Count	Proposed Labels as per *Dataset 1*
Traffic	63	B (45.4%)
Falls	236	
Fire and Explosion	47	
Electrocute	108	
Collapse of object	212	C (32.3%)
Caught in between	68	
Struck by falling objects	43	
Exposure to chemicals	29	D (4.6%)
Exposure to extreme temperature	17	
Others	43	-- (17.7%)
Struck by moving objects	134	

3.3 Classification of SOs with GLDA seeded using YAKE keywords

The probabilistic model underlying GLDA assumes that document sets can be divided into latent topics, and each topic is made up of different words [7]. It uses Dirichlet distributions in the form of document-topic distribution and topic-word distribution and identifies the topic or category a particular document belongs to, using an iterative procedure. GLDA model uses the domain-specific keywords extracted using YAKE for seeding purposes so that the words are not randomly assigned to a topic during initialization and the topics generated are human interpretable. For seeding, a non-zero weightage

is assigned to the domain-specific keywords during the initialization of GLDA. The *Seed Confidence* (SC) parameter of GLDA can control the weightage assigned to these seed keywords, ranging between 0 and 1. The overall implementation of GLDA is consistent with the previous work [12], which provides details on the involved mathematical formulations. GLDA then uses the YAKE generated seed keywords and the main SOs from the input dataset to generate document-topic distribution. The document-topic distribution provides the probabilities of a document belonging to each of the four categories. The SOs are labeled with the category showing the maximum probability. The classification performance is then evaluated using the commonly adopted F1 score metric [3,7], where the GLDA predicted category labels are compared with those present in the original datasets.

4 Results

4.1 Top-Keywords from YAKE

Table 3 presents an overview of the top-10 (based on estimated weights) keywords extracted for each article. Overall, the top keywords provide an intuitive validity of the keyword extraction process using YAKE. For example, keywords such as "glove," "ppe," "wear," and "protect" represent the category label A, i.e., PPE.

Table 3. Top-10 keywords obtained using YAKE for each category label

A	B	C	D
protection	safety	equipment	waste
glove	fire	conductor	material
ppe	equipment	part	construction
protective	risk	tool	recycling
wear	site	employee	demolition
protect	electrical	material	management
hazard	worker	ladder	project
equipment	ladder	volt	building
type	construction	metal	separate
safety	excavation	expose	product

4.2 F1 scores for *Dataset 1*

Table 4 summarizes the optimal F1 scores obtained for different category labels and the weighted F1 score. Optimal parameters obtained through a qualitative sensitivity analysis approach are also summarized in Table 4. The optimal F1 scores represent the maximum of 152 runs representing 2, 4, and 19 variations of hyperparameters "Alpha," "Iterations," and "SC." *Alpha* is the Dirichlet parameter for distribution over topics, while *Iterations* is the number of sampling iterations done by the LDA algorithm before convergence. The optimal total F1 score of 0.62 is obtained. The results also indicate a considerable variation in F1 scores for individual categories. The best performing category is "A," with an optimal F1 score of 0.801. The poorest performing category is "C," with an F1 score as low as 0.37 (see Table 4).

Table 4. F1 scores for Dataset 1

Category	F1 Score	Optimal Hyperparameters*
A	0.801	(0.01, 5000, 0.65)
B	0.56	(0.02, 4000, 0.98)
C	0.37	(0.02, 5000, 0.30)
D	0.69	(0.01, 5000, 0.55)
Total Score**	0.62	(0.02, 4000, 0.98)

*Optimal Hyperparameters are represented using the following notation – ("Alpha," "Iterations," "SC")
** For the maximum Total F1 score, the F1 score for each of the categories are – A (0.796), B (0.56), C (0.34), and D (0.69)

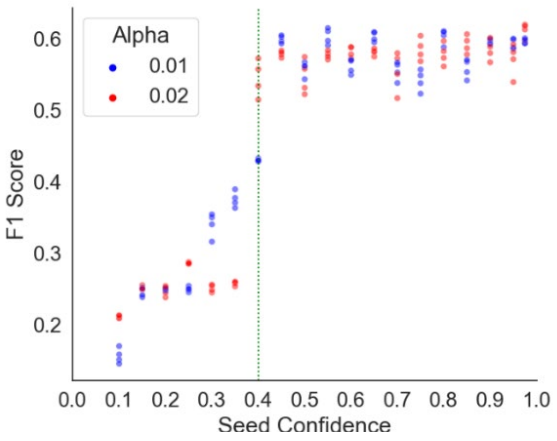

Figure 2. Total F1 Score sensitivity with "SC" and 'Alpha' parameters for *Dataset 1*

Figures 2 and 3 also summarize the results from sensitivity analysis for the total F1 score and individual category labels. In both figures, four distinct points corresponding to a specific alpha and SC show the different results in variations in the parameter "Iterations." Overall, parameters "Alpha" and "Iterations" do not have a high impact on the F1 score. On the other hand, very high sensitivity in the F1 score around a critical value of the hyperparameter "SC" is obtained. Results from Figures 2 and 3 also suggest that the optimal parameters are different for each category.

4.3 F1 scores for *Dataset 2*

Table 5 summarizes the optimal F1 scores obtained

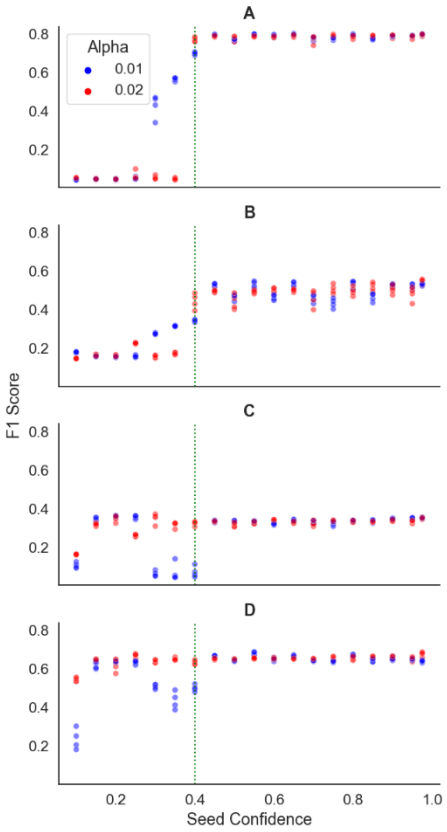

Figure 3. Category wise F1 Score sensitivity with "SC" and 'Alpha' parameters for *Dataset 1*

for different category labels and the weighted F1 score for *Dataset 2*. The optimal total F1 score of 0.48 is obtained. A considerable variation in F1 scores for individual categories is also observed. The best performing category is B, with an optimal F1 score of 0.62. The poorest performing category is D.

Table 5. F1 scores for Dataset 2

Category	F1 Score	Optimal Hyperparameters
B	0.62	(0.01, 5000, 0.30)
C	0.56	(0.01, 5000, 0.95)
D	0.24	(0.02, 2000, 0.60)
Total Score**	0.48	(0.01, 5000, 0.30)

** For the maximum Total F1 score, the F1 score for each of the categories are –B (0.62), C (0.33), and D (0.04)

5 Discussions

5.1 Added value of the YAKE-GLDA approach

Overall, a moderate to high classification performance of (F1 score of 0.62 for Total data and 0.8 for category A) has been obtained for the near-miss dataset representative of conditions at the real construction site using the YAKE-GLDA approach. Such classification performance is at par with the previously known application of YAKE-GLDA in the aviation industry [7]. Even for the construction industry, such classification performance is at par with previously reported classification performances obtained using various supervised learning approaches [3]. However, it is also important to note that very high levels of classification performance have also been reported in previous studies utilizing various supervised approaches [5]. Hence, there is still significant scope for further improving the classification performance using the YAKE-GLDA framework. However, the YAKE-GLDA approach has a significant advantage in reducing the need for manual work compared to supervised approaches. The automated keyword extraction process is highly efficient as quality information on various category labels can easily be obtained from domain corpora created using commonly available resources on the web. Furthermore, the GLDA does not require any pre-labeled dataset for learning and classification tasks. In contrast, a requirement for large quantities of the pre-labeled dataset is one of the most significant limitations for the practical implementation of the supervised approaches on construction sites [4,18].

Furthermore, a comparison of results obtained from *Dataset 1* and *Dataset 2* suggests a high degree of generalizability for the domain corpus to obtain classification performance across different types of datasets within the same industry. Typically, ML models trained for one set of datasets perform poorly on classification tasks for similar datasets from different sources [19]. Despite significant differences in the two datasets, domain corpus developed for one application can achieve high classification performance for another dataset, e.g., for categories B and C, where higher classification performance has been obtained than Dataset 1. Hence, such preliminary results of the YAKE-GLDA suggest a high potential for rapid and broader implementation of the framework across different construction sites.

5.2 Classification performance and ideas for its improvement

One of the significant aspects affecting the classification performance in the YAKE-GLDA approach is related to the characteristics of the input data in the construction sector. Classification labels typically used in the construction industry are rarely mutually exclusive, creating challenges for ML approaches to classify the observations in a single category [3]. The same is also observed in the current study. Table 6 shows each topic's top words as directly generated by the

GLDA's topic-word distribution based on *Dataset 1*. As highlighted in red in Table 6, many of the top words between categories "B" and "C" are common. Such commonality in the topic words can influence the GLDA's classification accuracy. For example, for the most frequent category in *Dataset 1*, i.e., "B," about 55% of the observations were classified incorrectly. About 50% of these incorrect classifications were classified in category "C." Many ideas on managing challenges related to characteristics of the input data can be implemented in the future studies. GLDA's ability to classify a single observation in multiple-categories should be explored. Furthermore, the unstructured input data need to be converted to structured data as much as possible. Hence, dense vector representation for the text, such as word embeddings, should be explored with GLDA to enhance the performance of the classification tasks[20].

Table 6. Top words for each category for *Dataset 1* as obtained through topic-word distribution in GLDA.

A	B	C	D
circulation	circulation	circulation	our
playful	kind	listed	kind
multiple	workingacceing	underground	underground
listed	towerlight	instruments	instruments
ice	companion	playful	note
earthing	clapboard	earthing	re
underground	toward	molten	inspection
barrication	underground	end	containment
goal	earmuf	electricityhot	fund
instruments	instruments	ice	onsiteafter
seat	earthing	earmuf	barricadding
child	playful	gloves	clapboard
blink	barricadding	barrication	molten
random	double	age	prepare
sss	smoking	kind	tables
workingacceing	fund	dressing	communicate
peak	our	crushed	greenfield
molten	foundation	pope	pepsi
mu	trans	cabin	collar
messy	begin	insulatedwhich	reddy

On the other hand, the domain corpus's quality and comprehensiveness are another significant aspects affecting the classification performance in the YAKE-GLDA approach [7]. Even in the current study, information for some of the subcategories for category C, such as the information related to color-coding of equipment, could not be readily obtained. Hence, the classification performance for category C for *Dataset 1* is inferior (F1 score of 0.37). Whereas, for *Dataset 2*, in which equipment color coding-related factors were not available, the classification performance for category C is significantly high. Hence, a focus of subsequent study could also be to enrich the domain corpus [7].

5.3 Study limitations

The current study is the first to explore the applicability of YAKE-GLDA approach for analyzing the construction SOs. Overall, several promising results have been obtained. However, there are significant limitations of the work requiring improvement.

The current study relies on qualitative methods to assure the validity of the domain corpus generation and data harmonization process. Even though the safety-related experience of authors has helped in the process, more rigorous validations relying on inputs from multiple domain experts are necessary.

The analysis in this study has been focused on classification tasks centered at somewhat broader category levels. However, construction organizations could benefit from tracking trends for other refined categories. In such conditions, the value addition of the YAKE-GLDA approach should also be demonstrated for classification performance at the micro subcategory levels. Such detailed categories are also available in Dataset 1 and could be explored in future studies.

The study's novel sensitivity analysis results also highlight a lack of generalizable trend in F1 scores with varying SC. Figures 2 and 3 indicate a steep improvement in the F1 score around a critical value of SC and a relatively invariable trend afterward. However, the optimal hyperparameter combination is also different for each category. Such a lack of generalizable trends in sensitivity to hyperparameters indicates potential consistency issues for the approach and may restrict its applicability for new unlabeled data. The results also underscore the importance of conducting a sensitivity analysis to obtain optimal values of the parameters to be used in all subsequent applications. Furthermore, more research is deemed necessary to fully understand the sensitivity of the approach on more parameters that could guide its applications to different construction sites.

6 Conclusions

The current study is the first application of an unsupervised YAKE-GLDA approach for the fully automatic classification of SOs in construction. The process reduces the necessity of manual intervention significantly, provides a moderate classification performance (F1 score of 0.81 for ppe category), and is potentially generalizable to different data sources related to safety in the construction sector. On the other hand, previously unexplored sensitivity to hyperparameters reveals a non-generalizable trend affecting YAKE-GLDA's new application to an unlabeled dataset. Hence, future research is vital for assuring the approach's applicability to various construction sites. Efforts are also necessary for improving the classification performance of the approach. Finally, the study's limitations should

also be addressed to make an objective assessment of the applicability of YAKE-GLDA approach for efficient analysis of SOs in construction.

References

[1] P. Manu, F. Emuze, T.A. Saurin, B.H.W. Hadikusumo, *Construction Health and Safety in Developing Countries*, Routledge, 2019.

[2] N. Bugalia, Y. Maemura, K. Ozawa, A system dynamics model for near-miss reporting in complex systems, *Saf. Sci.,* 142:105368, 2021. https://doi.org/10.1016/j.ssci.2021.105368.

[3] Y.M. Goh, C.U. Ubeynarayana, Construction accident narrative classification: An evaluation of text mining techniques, *Accid. Anal. Prev*., 108: 122-130, 2017. https://doi.org/10.1016/j.aap.2017.08.026.

[4] S. Sarkar, J. Maiti, Machine learning in occupational accident analysis: a review using science mapping approach with citation network analysis, *Saf. Sci.* 131:104900, 2020. https://doi.org/10.1016/j.ssci.2020.104900.

[5] S. Baek, W. Jung, S.H. Han, A critical review of text-based research in construction: Data source, analysis method, and implications, *Autom. Constr.* 132:103915, 2021. https://doi.org/10.1016/j.autcon.2021.103915.

[6] A. Chokor, H. Naganathan, W.K. Chong, M. El Asmar, Analyzing Arizona OSHA injury reports using unsupervised machine learning, *Procedia Eng.* 145, 1588–1593, 2016.

[7] A. Ahadh, G.V. Binish, R. Srinivasan, Text mining of accident reports using semi-supervised keyword extraction and topic modeling, *Process Saf. Environ. Prot.* 155: 455-465, 2021. https://doi.org/10.1016/j.psep.2021.09.022.

[8] C.-W. Cheng, C.-C. Lin, S.-S. Leu, Use of association rules to explore cause–effect relationships in occupational accidents in the Taiwan construction industry, *Saf. Sci.,* 48: 436-444, 2010. https://doi.org/10.1016/j.ssci.2009.12.005.

[9] M.N. Moreno, S. Segrera, V.F. López, Association Rules: Problems, solutions and new applications, *Actas Del III Taller Nac. Minería Datos y Aprendizaje*, Tamida, 317–323, 2005.

[10] A.J.-P. Tixier, M.R. Hallowell, B. Rajagopalan, D. Bowman, Automated content analysis for construction safety: A natural language processing system to extract precursors and outcomes from unstructured injury reports, *Autom. Constr.,* 62:45-56, 2016. https://doi.org/10.1016/j.autcon.2015.11.001.

[11] J. Jagarlamudi, H. Daumé III, R. Udupa, Incorporating lexical priors into topic models, in: *Proc. 13th Conf. Eur. Chapter Assoc. Comput. Linguist*., 204-213, 2012.

[12] S. Zhou, P. Kan, Q. Huang, J. Silbernagel, A guided latent Dirichlet allocation approach to investigate real-time latent topics of Twitter data during Hurricane Laura, *J. Inf. Sci.,* 2021. https://doi.org/10.1177/01655515211007724.

[13] Y. Suh, Sectoral patterns of accident process for occupational safety using narrative texts of OSHA database, *Saf. Sci.,* 142:105363, 2021. https://doi.org/10.1016/j.ssci.2021.105363.

[14] R. Wang, W. Liu, C. McDonald, Using word embeddings to enhance keyword identification for scientific publications, in: *Australas. Database Conf.*, Springer, 257-268, 2015.

[15] S.J. Rose, W.E. Cowley, V.L. Crow, N.O. Cramer, Rapid automatic keyword extraction for information retrieval and analysis, 2012.

[16] R. Campos, V. Mangaravite, A. Pasquali, A. Jorge, C. Nunes, A. Jatowt, YAKE! Keyword extraction from single documents using multiple local features, *Inf. Sci. (Ny)*. 509:257-289, 2020. https://doi.org/10.1016/j.ins.2019.09.013.

[17] N. Giarelis, N. Kanakaris, N. Karacapilidis, A Comparative Assessment of State-Of-The-Art Methods for Multilingual Unsupervised Keyphrase Extraction, in: *IFIP Int. Conf. Artif. Intell. Appl. Innov.*, Springer, 635–645, 2021.

[18] J. Kedia, T. Vurukuti, N. Bugalia, A. Mahalingam, Classification of safety observation reports from a construction site: An evaluation of text mining approaches, in: *PMI Res. Acad. Virtual Conf.* 50–66, Indian Institute of Technology Bombay, Mumbai, 2021.

[19] L. D'hooge, T. Wauters, B. Volckaert, F. De Turck, Inter-dataset generalization strength of supervised machine learning methods for intrusion detection, *J. Inf. Secur. Appl*. 54:102564,2020. https://doi.org/10.1016/j.jisa.2020.102564.

[20] F. Zhang, A hybrid structured deep neural network with Word2Vec for construction accident causes classification, *Int. J. Constr. Manag*. (2019) 1–21. https://doi.org/10.1080/15623599.2019.1683692.