A pre-trained language model-based framework for deduplication of construction safety newspaper articles

Abhipraay Nevatia[1], Soukarya Saha[2], Sundar Balarka Bhagavatula[3] and Nikhil Bugalia[3*]

[1] Department of Mechanical Engineering, Indian Institute of Technology Madras, India
[2] Department of Engineering Design, Indian Institute of Technology Madras, India
[3] Department of Civil Engineering, Indian Institute of Technology Madras, India
E-mail: me20b007@smail.iitm.ac.in, ed20b062@smail.iitm.ac.in, ce18b103@smail.iitm.ac.in, *nikhilbugalia@gmail.com

Abstract –

The unavailability of Occupational Health and Safety (OHS) statistics for the construction sector is a systemic hurdle in improving safety, particularly for developing countries. Alternatively, online newspaper articles are deemed a potential source for OHS statistics. Machine Learning (ML) approaches for text-mining are essential for the otherwise resource-intensive processing of news articles. However, the previous literature applying ML for newspaper articles has been scarce, and tasks, such as removing duplicate reports, have not been addressed satisfactorily. The current study develops and evaluates a novel framework based on pre-trained language models for the deduplication tasks for construction safety-related news articles to address the research gap. The study relies on the Question and Answering (QA) ability of the Longformer model pre-trained on Stanford QA Dataset (SQUAD) to identify the date and location of the construction accidents from the news articles. A combination of date and location is used as a key for deduplicating news articles that refer to the same accidents. The comparative performance of the developed framework is evaluated on 141 accident articles systematically extracted from 7 months of construction-relevant news articles in India. With an accuracy of more than 90%, the proposed method outperforms other methods for date identification. The performance of the deduplication process based on Longformer, i.e., F1 score of 0.79, is comparable to the Cosine similarity-based approaches. However, compared to the commonly adopted Cosine similarity-based method, the newly developed method in this study is reliable and consistent for periodically processing large quantities of unlabeled datasets.

Keywords –
Construction safety, News Articles, Machine Learning, BERT, Longformer, Deduplication

1 Introduction

For developing countries such as India, the construction sector remains one of the worst-performing sectors for Occupational Health and Safety (OHS) matters [1]. For policymakers wishing to solve the issue, one of the essential ideas to improve safety performance is to collect and analyze data on accidents, injuries, and near-miss reports and utilize the learning from these reports to enable sector-wide safety measures [2]. However, the unavailability of robust OHS statistics for the construction sector in developing countries is a systemic hurdle facing academia and practitioners [3], where government agencies have no formal mechanisms to collect and publish such statistics.

Without formal databases, online newspaper articles have been recognized as a potential source of information for developing OHS statistics [3,4]. However, previous attempts to leverage large-scale online news data for developing such safety statistics for the construction sector have been scarce, and the challenges faced in such analysis have not been well-addressed [4]. For example, only a handful of previous studies have addressed the resource-intensiveness-related problem related to processing large quantities of text data [5], as generally found in news articles [4]. A few of them have relied on efficient data processing approaches such as Machine Learning (ML) and text-mining for construction-related news items [6,7]. Even within these ML studies on construction news articles, issues such as identifying duplicate news articles have not been addressed appropriately. Current studies typically rely on text-similarity-based approaches to compare news articles and detect duplicates [6]. However, such text-similarity-based approaches lack consistency in creating a database of accident articles synthesized through large quantities of news articles that can be updated periodically (see section 2 for details).

The essential motivation for this paper is to develop a novel approach that can accurately and consistently

identify the duplicates in construction safety-related news articles. Consequently, the current study aims to develop and test a novel duplicate identification approach relying on state-of-the-art pre-trained language models, similar to Bidirectional Encoder Representations from Transformers (BERT). The study makes essential contributions to advancing the usage of ML approaches in developing reliable trends on OHS statistics using newspaper articles, especially in countries where industry-wide reporting on OHS data is non-existent.

The study is structured as follows. Section 2 provides an overview of the literature and identifies the essential gaps to where the study contributes. Section 3 describes the essentials of BERT-based language models and the analytical methodology adopted in the current study. Results have been summarized in section 4, followed by discussions in section 5. Conclusions have been outlined in section 6.

2 Literature Review

One of the most comprehensive analyses of fatal accidents in the construction sector using newspaper data has been presented in [4]. Using statistical approaches such as cluster analysis and principal component analysis, they obtained key statistics related to accidents, such as the date and time of the fatal accidents. Overall, the study provides comprehensive ideas on processing newspaper reports; however, the manual data collection and entity extraction processes adopted were resource intensive. Only a handful of previous studies have relied on automated ML and text-mining approaches to analyze newspaper data for construction safety. For example, Feng & Chen [8] proposed a natural language data augmentation-based framework for automatic information extraction using deep neural networks. However, the articles used for their study were handpicked, and their framework can only be applied to small datasets extracted manually. They emphasize the necessity of a robust automatic information extraction model capable of handling large volumes of data [8]. The challenge of automatically extracting large quantities of construction safety data from newspapers has been partially addressed in [9]. They relied on a keyword-based extraction technique to collect news articles. Their data helped identify factors and interrelationships affecting fire-related accidents in construction.

However, even Kim et al. [9] do not address some fundamental challenges facing the widespread usage of ML approaches for generating OHS statistics in the construction sector [6]. For example, the keyword-based extracted data on newspaper articles may contain duplicates. The duplicates are of different types. Examples include - the same accident being reported in print and online format, multiple media houses reporting the same construction accidents, and a piece of news referring to a previously reported construction safety event [6]. Removal of duplicates, also known as the deduplication process, is essential to avoid obtaining overestimated OHS statistics [4].

The commonly adopted deduplication approaches rely on using vector space techniques such as Cosine similarity, Jaccard similarity, and Euclidean distances to calculate the text similarity [10]. However, such models are inherently limited. Specifically, the news articles sparsely contain the target safety-relevant words, and most of the textual information is generic [6]. In such conditions, text-match-based scores may not be precise. One of the most significant limitations of vector space models for deduplication tasks is their lack of trainability and consistency. Vector space models make a pair-wise comparison of articles and generate a text-similarity score. Afterward, the analysts must set a threshold value, and articles above the threshold value are considered duplicates. Such a threshold value is optimized using an annotated dataset for identifying duplicates. However, Barbera et al. [6] note a significant variability in the data distribution obtained through keyword extraction-based approaches across different periods. Furthermore, it is implausible that two sub-sets of the same extensive data will follow a similar distribution. Such data limitations make it challenging to set the vector-space model's threshold values for deduplication tasks that can be consistently applied to the whole dataset. Hence, alternative ML and text-mining methods for the deduplication process must be explored. However, such explorations have been rare in the existing literature.

Sitas et al. [11] provided a framework for designing a deduplication process. They recommend identifying relevant "fields" in the data. Such *fields* should be relatively stable constructs despite the variability in the overall information in multiple records. A combination of multiple such fields can then be used to create "keys," which can be used for deduplication. For negative consequence-related information in news articles, the date of the event and the broad region or locality of the adverse event is most-commonly described. Hence, the accident's date and location are potential candidates to be used as *fields*, and a combination of them can serve as a potential *key* for the deduplication process [12].

For extracting the date and location-related information from the textual data, many previous studies have utilized different text-mining approaches broadly under the category of tasks known as Named Entity Recognition (NER). The NER tasks are either Rule-Based approaches or Supervised ML approaches. However, all the above methods are resource-intensive, requiring manual efforts to identify patterns or annotate the data. On the other hand, natural language processing has witnessed a significant paradigm shift since the

introduction of the BERT method. BERT is a language model pre-trained on large volumes of unlabelled texts. BERT representations can achieve state-of-the-art results on several language processing tasks. For example, recent studies have also shown that BERT based model can outperform conventional methods even in the NER tasks [13]. While such work is commendable, the tagged text phrases still need to be manually interpreted to identify the date and the location of a construction-related accident, as often there can be mentions of multiple dates and places in a single news article. In recent years, significant progress has been made in the BERT-based models across various language processing-related tasks such as summarization and question-answering [14]. Some of these advancements have the potential to make the overall process of date and location identification in newspaper articles efficient and less resource intensive.

However, to the authors' knowledge, none of the existing studies have utilized the BERT-based language models to analyze construction safety-related news articles. An efficient and resource-effective process for deduplication based on BERT will significantly advance the body of knowledge. It will pave the way for leveraging readily available news articles for developing safety-related OHS statistics.

3 Methodology

The overall analytical process adopted for the current study has been summarized in Figure 1. The focus of the analysis has been to understand the comparative performance of the BERT-like pre-trained language models in deduplication tasks with conventional text-similarity-based methods and manual estimation.

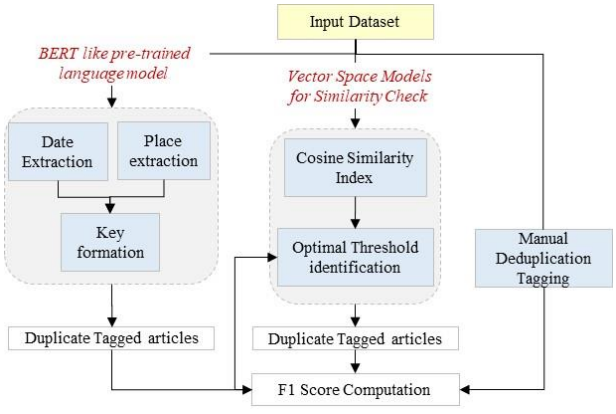

Figure 1. The analytical process of the study

3.1 Input Dataset and Manual Deduplication

The current study relied on the services of a news media analytics company, which has access to digital news articles from all major news agencies in India to obtain data. Consistent with the recommendations of the literature, a keyword-based search was used to extract relevant news articles. The keywords used for extraction are 'Construction', 'Accident', 'Injur*', 'Fall', 'Collapse', 'Struck', 'Dead', 'Worker'. The 'Construction' keyword is mandatory in the given keywords. This way, 11,208 articles reported between July 2021 and January 2022 were obtained. Multiple safety experts manually examined the text of each article to identify the articles describing construction accidents. Despite such an extensive manual effort, only 141 articles (1.26% of the total) contained information on actual construction accidents.

Three authors then implemented a pair-wise comparison scheme to tag the duplicate articles. Such information was stored in a 141*141 matrix, where the cell (i,j) was marked as "1" if the article in i^{th} row was found to duplicate the article on the j^{th} column. Otherwise, a value of "0" was assigned to cell (i,j). The manual deduplication matrix thus created is the *Truth Matrix* and has been used to evaluate the performance of the other algorithms. Such a process revealed the significant extent of deduplication in the news articles, where 77 out of 141 articles had at least one duplicate. On the other hand, 1 single accident also matched with 26 other articles.

3.2 Vector-Space model for similarity check

Consistent with the previous literature, the current study also develops a Cosine similarity-based text-match score to estimate the duplicates. The purpose of including the results from the Vector-space model is to provide a comparative assessment of the proposed model. Like the steps described above for the *Truth Matrix*, a 141*141 matrix containing a cosine similarity score from a pair-wise comparison of the articles is first developed. Then, a threshold value is selected, and any score above the threshold represents duplication in the articles and is marked as "1". Similarly, scores below the threshold are marked as "0".

The optimal cosine similarity threshold is identified through a sensitivity analysis approach. The results between the *Truth Matrix* and the cosine-similarity matrix are compared for generating F1 statistics for any arbitrarily selected threshold value. F1 statistics is a common approach to evaluate the various ML algorithm's performances compared to the truth data on binary classification tasks [1]. F1 score ranges between 0 and 1, and a higher F1 score represents a better performance for any algorithm. The threshold values are identified such that the F1 score is maximized.

3.3 Pre-trained language model-based novel deduplication framework

Consistent with the recommendations from the

previous work on deduplication, accident date, and location are considered the two *fields* commonly reported in news articles. Combining the two *fields* can be a unique *key* to help identify duplicate articles [12].

3.3.1 Date identification

Transformer-based pre-trained language models like BERT are among the most significant breakthroughs in text-mining and natural language processing. However, even BERT cannot process long text sequences (more than 512 tokens). Longformer has been developed as an alternative. Longformer's architecture is like BERT's architecture but has a different self-attention mechanism [14]. Literature has also shown that Longformer outperforms the BERT-based models for long text sequences in NER tasks [14]. Many news articles obtained in the current study are longer than 512 tokens; hence, Longformer implementation is deemed more suitable.

The Longformer model is primarily efficient in Question and Answering (QA) tasks [14]. The QA ability of the Longformer allows the user to ask questions in natural language, for which answers are sought from the target article, even if the exact sequence of the words present in the questions is not present in the target text being searched [14]. Considering these advantages, the current study uses Longformer for QA pre-trained on Stanford QA Dataset (SQUAD), mainly comprising of information from Wikipedia [15]. The QA query aimed to get an output that could readily help estimate the date of the accident. The exact question has been fine-tuned with several trials. Once the news articles are parsed through the Longformer, the output is a string. The information in this string is converted to an exact date using a reference for the *article publishing date* mentioned in the online article. The *Datetime* library in Python is used for such implementation. The *Datetime* library recognizes the "on <weekday>" format and not just "<weekday>." Therefore, in the *Datetime* function, the term "on" has been inserted before the <weekday>. The accuracy of the Longformer formulation is measured by comparing the manually extracted dates. In some instances, such a process also returns a NULL value.

3.3.2 Location identification

Like the date identification task described above, the current study uses Longformer for location identification tasks. However, in many cases, Longformer's output was insufficient to identify the accidents' location suitably. Hence, additional efforts for location identification are needed. Another pre-trained NER algorithm, Locationtagger, can identify all the places in the articles. Further, even when NER models are ineffective, the whole article was parsed as a string. Combined output is obtained by concatenating the output of each of the three methods to detect location, i.e., Longformer, NER-tagger, and the whole article parsing (See Figure 2). Each term in this combined string is then compared against a database containing names of Indian cities and places. The first term in the concatenated string that exactly matches with the database is estimated as the location of the accident. In such a manner, the highest priority is given to the location identified through Longformer, as the method is expected to capture the context of the description very well. The process for location identification is also summarized in Figure 2. The accuracy of the above formulation is measured by comparing the location of the accident manually extracted by the authors.

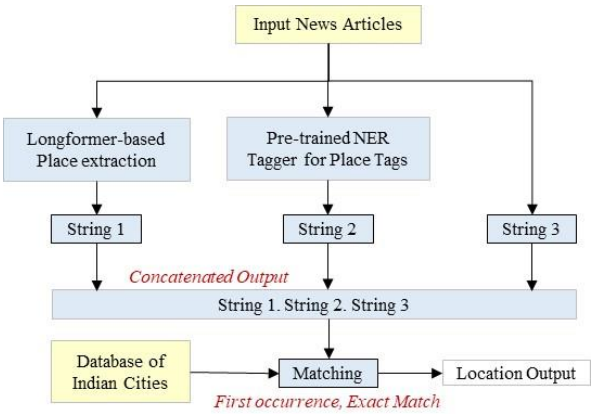

Figure 2. Location identification process

3.3.3 Key Formation

The output of the date and the location identification is concatenated to develop a *Key* that can serve as a unique identifier for a construction accident reported across multiple news reports. However, this approach's efficacy depends on the cases when both date and location are suitably identified from a news report. However, a NULL output may sometimes be possible from either date or location identification processes. Hence, two scenarios have been conceptualized to incorporate the possible NULL output for developing the *Key*.

Scenario 1: If either date or place is NULL, the study assumes the article is unique and does not compare it with any other article.

Scenario 2: If either date or place is NULL, the study removes the article from its dataset and further comparison.

Similar to the approaches defined above, a pair-wise comparison matrix is developed to predict two articles as duplicates when their *Key* matches (marked as 1). If the *Key* does not match, the articles are predicted as unique (marked as 0). The matrix thus generated is then compared with the *Truth Matrix* to compute the F1 score, an indicator of the deduplication process's efficiency.

3.4 Implementation

All algorithms have been developed using the Python programming language. The Longformer model pre-trained on SQUAD dataset is readily available in the *Huggingface* library of Python and does not require any other modifications for processing news articles. The reference library and Longformer implementation can be referred to from the following LINK.

4 Results

4.1 Accuracy of the date identification

Table 1 shows the results of different questions asked to the Longformer model and the corresponding accuracy in identifying dates compared to manually extracted true dates. The results demonstrate that the QA method for Longformer is highly efficient and can successfully identify the date of accidents in 124 articles out of 129 articles where the accident date was present in the original text. Similarly, the Longformer correctly returns a NULL output in 8 out of 12 cases where the accident date was absent in the original data. Although the results from the QA model are also a function of the QA query, which requires fine-tuning is essential to get good results.

Table 1. Date identification accuracies for Longformer

QA for Longformer	Number of articles Original Data (Correctly Predicted by the Longformer)	
	With Dates	With No Dates
"What day of the week did the accident take place"	129 (124)	12 (8)
QA – "When did the accident take place", returned the time of the accident if the time is mentioned. But is not able to detect the day in many cases. Correctly predicts the results in 82/141 cases. QA – "What Day did the accident take place". Returns the time in some cases. Correctly predicts the 117 instances out of the total 141.		

4.2 Accuracy of Location Identification

For the query "*Which city did the accident take place in*" the Longformer-based QA method can correctly identify the locations in only 62 out of 141 articles. The results demonstrate that identifying the location in the news articles is challenging. The accuracy of location identification improved significantly when the Longformer's output was further combined with the output of the other models, as shown in Figure 2—such combined model results in the correct location prediction for 112 articles out of 141.

4.3 Deduplication Results

Table 2. Comparative performance of various methods on Deduplication efficiency

Results	Scenario 1	Scenario 2
Key (Date and Location), Location – Combined Model		
Null articles Removed	-	27
Accuracy	0.97	0.99
Precision	0.95	0.95
Recall	0.34	0.68
F1 Score	0.50	0.79
F1 Score (Cosine)	0.72	0.80
Threshold (Cosine)	0.20	0.24
Key (only date)		
Null articles Removed	-	12
Accuracy	0.98	0.99
Precision	0.91	0.91
Recall	0.68	0.88
F1 Score	0.78	0.90
F1 Score (Cosine)	0.72	0.77
Threshold (Cosine)	0.20	0.21

Table 2 shows the comparative analysis of deduplication efficiency between the proposed Longformer-based model and the Cosine similarity. Comparative analysis is also performed for both Scenario 1 and Scenario 2. Results from two different types of Keys have also been shown in Tables 2 and 3. The first Key relies on both fields, i.e., date and location. The second Key relied only on the date as the *field*. For comparison, results from Cosine similarity and the optimal cosine thresholds have also been included in Table 2. The results thus obtained have been discussed in detail in the next section.

Table 3. Confusion Matrices for various methods and scenarios

Results	TP	FP	FN	TN
Key (Date and Location), Location – Combined Model				
Longformer Scenario 1	161	8	311	9390
Longformer Scenario 2	161	8	77	6195
Cosine Scenario 1	325	105	147	9293
Cosine Scenario 2	181	31	57	6172
Key (only date)				

Longformer Scenario 1	319	30	153	9368
Longformer Scenario 2	319	30	43	7864
Cosine Scenario 1	325	105	147	9293
Cosine Scenario 2	268	68	94	7826
TP – True Positives, FP – False Positives, FN – False Negatives, TN – True Negatives. The number represents the count of duplicates (1) and non-duplicates (0) in the matrix.				

5 Discussions

5.1 Effectiveness of Longformer model in Date and Location extraction

Long-term spatial and temporal trends for OHS issues are essential in policymakers' decision-making. Hence, automatically identifying the date and the location of a given accident from newspaper articles is also a crucial problem to be solved. The pre-trained Longformer models adopted in the current study have shown promising date and location identification results. Notably, for date identification, the Longformer model can achieve an accuracy of 90% and outperforms the contemporary NER models on date identification by a significant margin [16]. The effectiveness of the Longformer model is further demonstrated by its ability to provide the correct accident date for a complex example of a news article shown below.

"The contractor's failure to follow safety measures led to the death of a worker engaged in the lifting of a concrete girder for the elevated highway on New Natham Road here ***on Saturday****, Public Works Minister E.N. Velu said* ***on Sunday****. After inspecting the site along with Finance Minister Palanivel Thiaga Rajan, Mr. Velu said a team of experts from the National Institute of Technology, Tiruchi, would hold an inquiry. ……. The construction of the elevated highway, stretching 7.3 km from Chokkikulam to Chettikulam, began* ***in November 2018****."*

In the above example, three instances that could indicate the date of an event are present (indicated in bold). One of these dates is related to the occurrence of the accident. The other date refers to a statement made by the regulatory authority on the accident. Nevertheless, another date refers to the event when the structure's construction began. The Longformer-based date detection model can distinguish these three dates and provide a correct output. Such a contextual interpretation of the dates is generally impossible, even for the pre-trained NER models relying on supervised ML approaches.

However, despite the excellent performance in date detection, even a Longformer-based formulation faces challenges in location identification. A quick analysis of the prediction errors provides an overview of some common reasons for poor location prediction performance. One of the primary reasons is that often, in news articles, the locations mentioned refer to streets, names of the localities, or residential colonies within a city. However, the database for comparison only contains information on the city. In such cases, a possibility of an exact match between the Longformer prediction and the matching database reduces significantly, leading to an overall poor performance of location prediction. Such error analysis also provides essential ideas for further improving the algorithm's performance on location identification. These ideas have been summarized in section 5.3 of the current study.

5.2 Effectiveness of Longformer-based deduplication process

Overall, several advantages of the Longformer-based deduplication process can be observed from the results obtained in the current study. The foremost advantage of the Longformer-based model, compared to the conventional Cosine similarity-based process, is the expected generalizability of the method. As shown in Table 3, the optimal threshold value for Cosine similarity can change significantly even for a subset of the same dataset. Results in Table 3 demonstrate how different the number of True Positives is when the Cosine similarity method is applied to a smaller subset of the larger dataset. Because of such fluctuations in the results of the cosine similarity, it cannot be considered a reliable method for the deduplication process for large quantities of unlabeled data. However, such challenges are absent in the Longformer-based pre-trained language models used for deduplication. The models are robust (even for a smaller subset, as shown in Table 3) and can be readily used consistently without retraining across different prediction issues.

Nonetheless, the current study identifies many areas where even the proposed Longformer-based model requires improvement. First, the Cosine similarity-based method still outperforms the novel Longformer-based model for deduplication tasks. In many cases where either the date and location are not present in the online article or the Longformer face difficulties in detecting the date and the location, the proposed Longformer-based approach suffers. For such cases, the efficiency of the Cosine similarity method for the deduplication process is significantly better (See Scenario 1 results in Table 2). In other cases (See Scenario 2 results in Table 2), the performance of the Longformer is like the Cosine similarity-based methods. However, in such cases, a few articles have been excluded from being considered in the deduplication process. At the same time, the cosine

similarity-based method can predict all articles.

The result from the study also demonstrates that the efficiency of the Longformer-based deduplication process is significantly hampered due to the combination of multiple *fields*. For the sample of news articles considered in the current study, a "date" only *key* leads to a significant improvement in the efficiency of the deduplication process is observed. Such results indicate that stricter criteria requiring exact matches between multiple *fields* to detect duplicates reduce the effectiveness of the duplication process. Hence, the study results would indicate that a straightforward solution could be to reduce the number of *fields* included in the deduplication process. However, such a solution may not be human-intuitive. For example, despite the enhanced performance of the only date-based deduplication process, it is not practical to expect that only one accident is expected in one day for a construction industry as large as it is in India.

5.3 Ideas for improving the performance of the deduplication process

The comparative assessment of the two methods highlights the complexities of the deduplication tasks for news articles. The conventional Cosine similarity-based deduplication process is simple and can be done on the whole corpus of the dataset. However, the approach is unreliable, especially when a large quantity of unlabeled data must be checked for duplication periodically. In contrast, the Longformer-based pre-trained language model is potentially generic and can work on the unlabeled dataset. However, its limitations in identifying dates and locations based on exact match criteria for each article pair pose significant challenges for their usage in the deduplication process. The Longformer method can be improved in two significant directions to solve the abovementioned challenges.

First, efforts should be made to enhance the Longformer's prediction efficiency for particular *fields*. The output from multiple QA queries can be combined to identify the date and location. Similar efforts could also be made to enhance the Longformer's predictive capability of accident location at a city level. For example, Longformer's model can be assessed to identify the text phrases which contribute most towards the outcome of the Longformer, i.e., access the Longformer's attention matrix. The NER-based search for locations only on these high-attention text phrases can potentially improve the prediction capabilities of the Longformer for locations.

Second, efforts could also be made to shift from an "exact match" paradigm to a "partial match" paradigm. For example, an approximate search for accident locations at a city level can also be done by correlating the multiple locations mentioned in the article. The article's publishing location could also be used as a reference point. However, such an effort will require understanding the spatial information-based libraries specifically for India. Further, the partial match paradigm could also be extended to develop a clustering approach based deduplication process. For example, Longformer can further extract more features about the accident, such as the number of people affected and the gender of those affected. Type of construction activities where accidents were reported, such as bridges, roads, and buildings, or type of construction accidents, such as falls from a height, among others, could also serve as features. Then various clustering techniques that can simultaneously leverage the multiple features could be used to identify duplicate articles. The clustering techniques are versatile and can also be trained therefore maintaining the generalizability benefits of the Longformer-based models compared to the Cosine similarity methods.

5.4 Study limitations and future scope

The advantages of the Longformer-based deduplication process developed in this study have been shown only for a small dataset of 141 news articles. The availability of such a small sample set was not intentional but was due to the sparseness of the construction accident news in the whole dataset searched for relevant keywords. Although the preliminary results from handpicked news articles from the USA processed through the proposed method show a good performance in date, place, and duplication identification (see Link). However, the generalizability of the method should also be explored for large datasets of articles comprising multiple geographic regions. Despite the advances shown in the paper here, it is expected that in a multi-lingual country such as India, only a fraction of the construction accidents will feature in English news articles. India features more than 20 major vernacular languages. The scope of the analysis should be extended to include articles from these vernacular languages in future studies. In principle, the BERT-based models efficiently develop work processes that simultaneously assimilate data from multiple languages. However, the deduplication tasks for information featured in multiple languages will be an academically exciting extension of the current work.

6 Conclusions

The current study develops and evaluates a novel framework based on pre-trained language models for deduplication tasks for construction-related news articles. The study relies on the QA ability of the Longformer model pre-trained on SQUAD to identify the date and location of the construction accidents from the news articles. A combination of date and location is used as a key to detecting duplicate news articles that refer to the

same accidents featured in multiple news reports. The proposed method outperforms other methods by correctly identifying the date of accidents in more than 90% of the articles. Although, detecting the location of the accident through Longformer continues to be challenging. Overall, the Longformer-based model outperforms the traditional Cosine similarity-based method in the deduplication tasks when only accident date is used as a key. However, for a more realistic *Key* involving date and location, the Longformer's performance is comparable to the Cosine similarity-based method. The foremost advantage of the Longformer-based model, compared to the conventional Cosine similarity-based models, is the expected generalizability of the method. The prediction based on Longformer models are robust (even for a smaller subset, as shown in Table 3) and can be readily used consistently across different prediction issues. The study contributes to the scarce body of knowledge on scarce ML applications for analyzing construction safety statistics using newspaper articles. The study's findings pave the way to automate the process of extracting and processing large quantities of news articles and use them to prepare reliable trends in OHS statistics. The unavailability of OHS statistics for the construction sector is a systemic hurdle in improving safety, particularly for developing countries.

References

[1] N. Bugalia, V. Tarani, J. Kedia, H. Gadekar, Machine learning-based automated classification of worker-reported safety reports in construction, *Journal of Information Technology in Construction*, 27:926–950, 2022. https://doi.org/10.36680/j.itcon.2022.045.

[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] D.A. Patel, K.N. Jha, An estimate of fatal accidents in Indian construction, in: *Proceedings of the 32nd Annual ARCOM Conference*, 1: 577–586, 2016.

[4] Y.-H. Chiang, F.K.-W. Wong, S. Liang, Fatal construction accidents in Hong Kong, *J Constr Eng Manag*, 144: 4017121, 2018. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001433.

[5] H. Gadekar, N. Bugalia, Automatic classification of construction safety reports using semi-supervised YAKE-Guided LDA approach, *Advanced Engineering Informatics*, 56: 101929, 2023. https://doi.org/10.1016/j.aei.2023.101929.

[6] P. Barberá, A.E. Boydstun, S. Linn, R. McMahon, J. Nagler, Automated text classification of news articles: A practical guide, *Political Analysis*, 29:19–42, 2021.

[7] J. Ninan, Construction safety in media: an overview of its interpretation and strategic use, *International Journal of Construction Management*, 1–9, 2021. https://doi.org/10.1080/15623599.2021.1946898.

[8] D. Feng, H. Chen, A small samples training framework for deep Learning-based automatic information extraction: Case study of construction accident news reports analysis, *Advanced Engineering Informatics,* 47: 101256, 2021. https://doi.org/10.1016/j.aei.2021.101256.

[9] J. Kim, S. Youm, Y. Shan, J. Kim, Analysis of Fire Accident Factors on Construction Sites Using Web Crawling and Deep Learning Approach, *Sustainability,* 13:11694, 2021. https://doi.org/10.3390/su132111694.

[10] R. Singh, S. Singh, Text similarity measures in news articles by vector space model using NLP, *Journal of The Institution of Engineers (India): Series B*, 102: 329–338, 2021. https://doi.org/10.1007/s40031-020-00501-5.

[11] A. Sitas, S. Kapidakis, Duplicate detection algorithms of bibliographic descriptions, *Library Hi Tech*., 26:287–301, 2008. https://doi.org/10.1108/07378830810880379.

[12] A. Abid, W. Ali, M.S. Farooq, U. Farooq, N.S. Khan, K. Abid, Semi-Automatic Classification and Duplicate Detection From Human Loss News Corpus, *IEEE Access*, 8:97737–97747, 2020. https://doi.org/10.1109/ACCESS.2020.2995789.

[13] S. Almasian, D. Aumiller, M. Gertz, BERT got a Date: Introducing Transformers to Temporal Tagging, ArXiv Preprint ArXiv:2109.14927, 2021.

[14] I. Beltagy, M.E. Peters, A. Cohan, Longformer: The long-document transformer, ArXiv Preprint ArXiv:2004.05150, 2020.

[15] P. Rajpurkar, J. Zhang, K. Lopyrev, P. Liang, Squad: 100,000+ questions for machine comprehension of text, ArXiv Preprint ArXiv:1606.05250, 2016.

[16] I. Jayaweera, C. Sajeewa, S. Liyanage, T. Wijewardane, I. Perera, A. Wijayasiri, Crime analytics: Analysis of crimes through newspaper articles, *in: 2015 Moratuwa Engineering Research Conference*, IEEE, 277–282, 2015. 10.1109/MERCon.2015.7112359.