IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 21 March 2023, accepted 28 April 2023, date of publication 2 May 2023, date of current version 10 May 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3272610

APPLIED RESEARCH

New Avenues for Automated Railway Safety Information Processing in Enterprise Architecture: An NLP Approach

ABDUL WAHAB QURASHI[1], ZOHAIB A. FARHAT[2], VIOLETA HOLMES[3], AND ANJU P. JOHNSON[1]

[1]School of Computing and Engineering, University of Huddersfield, HD1 3DH Huddersfield, U.K.
[2]Advanced Manufacturing Research Centre (North West), University of Sheffield, BB2 7HP Blackburn, U.K.
[3]School of Computing and Engineering, University of Huddersfield, HD1 3DH Huddersfield, U.K. *(Deceased)*

Corresponding authors: Abdul Wahab Qurashi (abdul.qurashi@hud.ac.uk) and Anju P. Johnson (a.johnson@hud.ac.uk)

This work was supported by the Institute of Railway Research, University of Huddersfield, and the Rail Safety and Standards Board (RSSB).

ABSTRACT Enterprise Architecture (EA) is crucial in any organisation as it defines the basic building blocks of a business. It is typically presented as a set of documents that help all departments understand the business model. In EA, safety documents are used to manage and understand safety risks. A novel similarity system for railway safety document processing is presented in this work. It measures the feasibility of automated updating of EA models with the Rule Book by verifying whether Rail Safety and Standards Board (RSSB's) Rule Book clauses are present and complete in existing EA models. Additionally, a Natural Language Processing (NLP) based search feature was developed to drill through the database to find similar existing rules, principles, and clauses based on semantic similarity. The result will display the most similar clauses and rules with similarity scores and document names. In this study, different pre-trained Electra Small, DistilBERT (Distillation Bidirectional Encoder Representations from Transformers) Base and BERT (Bidirectional Encoder Representations from Transformers) Base were used to embed text. Additionally, the similarity between document rules was measured by cosine similarity metrics. With conclusive evidence, our findings show that BERT Base exceeds the other embedding methods in the semantic comparison of documents.

INDEX TERMS Natural language processing, enterprise architecture models, distillation bidirectional encoder representations from transformers, cosine similarity.

I. INTRODUCTION

An enterprise architecture (EA) is a method of understanding an organization's future road map, business implementation plans, flows of information and technological capabilities [1]. Developing current, and future versions of this integrated view can assist enterprises in transitioning from current to future operating states [2]. Core elements of enterprise architecture consist of frameworks, standards, best practices, methodology and artefacts. Specifically, EA documents for railways include railway safety standards for ensuring the safety of complex interactive systems [3], risk assessment for railway departure process [4], and risk and safety decision-making for railway [5].

Safety documents are of key importance in the railway industry. These documents contain important information regarding railway safety and security, including manuals, operating procedures, and guidance documents. Machine learning (ML) and deep learning (DL) algorithms are used to digitalise the railway industry for safety risk assessment [6], [7]. In enterprise, vector models are often used to examine document similarity with previous rules, contracts and documents. This approach to NLP makes it possible to rank target documents or rules according to their similarity with a source

The associate editor coordinating the review of this manuscript and approving it for publication was Justin Zhang.

document which is an indicator of the relevance of the target document [8].

NLP is a subfield of linguistics and Artificial Intelligence that gives machines the capability of understanding and interpreting human language [9]. It allows machines to communicate with humans using human language. It also allows machines to read colossal text corpus and speech data and interpret it. NLP mainly bridges the communication gap between humans and computers [10]. Word embedding is a method of converting text in the form of real-valued vectors since machines are only capable of understanding numbers [11]. One-hot encoding is a way to encode words into numbers/vectors based on categorical features. A unique vector is assigned to every word with a length equal to the size of the dataset. The dimensionality problem prevents this approach from being used on large dictionaries [12]. Term Frequency-Inverse Document Frequency (TF-IDF) determine the importance of a word in a document. It is mainly used in document search, information retrieval, text summarization, and keyword extraction. However, it does not provide information regarding the similarity of documents [13]. Word2vec was introduced by Mikolov as a model for word or text representation in a vectorized format to understand the context and interpret it for NLP tasks [14]. Skip-gram neural network model is an unsupervised learning technique. When a pair of sentences is given as input, it predicts the surrounding words in the sentence. In contrast, Continuous Bag of Words (CBOW) predicts the target words based on the context words [15]. The word embedding based on convolutional neural networks shows improved results in measuring sentence similarity [16].

A Recurrent Neural Networks (RNNs) is an artificial neural networks model that utilizes previous outputs as inputs while maintaining a hidden state [17]. RNNs are more suitable for NLP applications, particularly Long Short Term Memory (LSTM) models that show effective results for sentiment classification [18] and Manhattan LSTM model for language translation tasks [19]. Bi-directional Long and Short Term Memory (BiLSTM) is another type of RNN that uses long sequence information to measure the semantic connection between words. The model consists of two warped LSTM layers, one receiving forward input and the other backward that helps it learn features from the adjacent layers [20]. The inherent sequential nature of the recurrent model makes it difficult to parallelize the long input sequence. A breakthrough in NLP took place in 2017 when transformers were introduced, which can handle varied input sequences. To compute representations of input and output sequences, the transformers use self-attention instead of RNNs or convolutions [21].

This paper developed a novel railway document safety model for semantically comparing two documents for railway safety and implementing a semantic search feature to match existing rules or clauses in both documents. Various aspects of safety are described in each document, including rules, objects, and responsible individuals. Pre-trained models such as Electra Small, DistilBERT Base and BERT Base were used to embed the document's rules. The similarity between the rules of the documents is analyzed by cosine similarity. In addition, a search-based system is devised that allows users to search for any rule, principle, object, or responsible actor across both documents.

A. MOTIVATION

In recent years, data has grown exponentially in railways due to digital transformation. Safety-related documents are of primary importance in the railway industry. Specifically, document similarity plays a critical role in designing a digital solution for railway document safety systems. Several documents in railway Enterprise Architecture cover safety aspects, including rules, principles and responsible actors. It is often necessary to consolidate multiple versions of the documents. As a result, it will be easier to provide cohesive advice and procedures in accordance with the same operational safety principles. In addition, these manuals and operational safety documents require analysis and processing to prevent the addition of redundant and duplicate rules.

It is possible to identify the similarity between different documents with the help of an expert that understand the context and semantic meaning of the rules present in multiple documents. This study measures the semantic similarity of two documents: the Enterprise Architecture (EA) Model and the Operational Concept Document (OCD). Additionally, a search-based system was developed to allow users to search for any rule, principle, or responsible actor in both documents.

B. CONTRIBUTION

A novel system was designed to process safety-critical railway documents. This method allows an expert to scan the whole procedure and verify the presence of rules between the two documents. If a new rule, object or responsible person is required, it can be easily searched through both documents to identify its existence to avoid duplication. By expert knowledge and hands-on identification, it was identified that the rule in the EA model document is brief compared to the OCD document. The system will also identify rules in different documents with the same semantic context. The main contribution of this paper is as follows:

- Developed a novel document processing system for railway safety-critical documents.
- Semantically comparing EA and RSSB rulebook documents for railway safety-critical systems to verify the presence of rules or clauses.
- A search-based system is devised that allows users to search for any rule, principle, object, or responsible actor across both documents.
- Compare the performance of pre-trained models such as Electra Small, DistilBERT Base and BERT Base for document processing.

- The similarity between the rules of the documents are analyzed by the cosine similarity metrics algorithm.
- Visualized the most frequent words in documents using the WordCloud library that provide insights and key trends for railway documents.

C. PAPER ORGANIZATION

The rest of the paper is structured as follows: Section II presents the literature review, followed by the proposed methodology for railway document processing in Section III. Section IV and V elaborate NLP tools and datasets details. Section VI presents the different sentence embedding techniques. Section VII discusses the text similarity metrics model, and Section VIII shows the pre-processing of data, document processing and evaluation of experimental results. Section IX and X conclude and provide direction for future work.

II. LITERATURE REVIEW

A. DOCUMENT PROCESSING FOR RAILWAY SAFETY ANALYSIS

The railway industry is a complex business that generates an enormous amount of data from multiple sources such as trains, tracks, employees, passengers, regulators and signalling systems. The documentation is constantly updated with multiple railway safety standards in place to guide the development and assessment of safety-critical software for railway control and protection systems. There are several safety standards that are commonly used in the railway industry such as IEEE 1474 [22], IEC 62278 [23], EN 50126 [24], EN50128 [25] and EN50129 [26]. These standards provide guidelines for risk evaluation, hazard identification, signalling, control and safety and protection management, as well as specifying requirements for safety-related electronic systems used in railway applications. In addition to the safety standards, there are also safety management documents, work recordings, maps, and operations data that are produced by railway organizations [27]. Each railway organization also develops its own enterprise architecture with roles and responsibilities. This ensures that everyone within the organization knows their role and responsibility in ensuring the safety of the railway system.

Overall, the railway industry is a complex business that generates a significant amount of data and documentation. The use of safety standards, safety management documents, and enterprise architecture helps to ensure that the railway system operates safely and effectively.

In railway, data produced from different sources is in an unstructured format and requires pre-processing and detailed analysis to provide context and key information. Multiple studies and surveys have been conducted to analyse this unstructured data in the railway industry. It was applied to Istanbul's automated fare collection system to provide better price recommendations to consumers for the BRT-Bus Rapid Transit line. It has also proposed recommendations for planning and management along with the graphical representation to provide insightful information [28]. In addition, railroad assets are also a great source of information that can be used for analyses and operational maintenance [29]. A multi-modal transport network was introduced in London based on the Markov chain approach. It was developed to handle complex data and provide better information [30].

Several NLP techniques have been successfully applied to document processing applications over the past several decades, including semantic analysis, language translation, text classification and summarization. Until recently, NLP systems relied on manually designed ontology rules. Due to the advancement, machine learning enabled the development of improved models as data volumes grew, which took advantage of ever-increasing amounts of information. The vector approach is prevalent in business applications, and today GPUs provide computational power sufficient for solving complex NLP tasks [31]. The University of Huddersfield has developed an ontology-based approach for extracting safety learning from a free text that captures human knowledge about real-world events in a knowledge model (an ontology) and is used to query documents [32].

Cluster analysis [33] and visual analytics [34] techniques are utilized to enhance the capability and performance of railway technical systems. The Rail Accident Investigation Branch (RAIB) is also applying NLP techniques to investigate the presence of failure-related entities from unstructured data [35]. Named Entity Recognition (NER) relies on a hybrid system combining Bayesian Learning and Conditional Random Fields (CRF) [36]. To measure text similarity and identify names and entities from unstructured text, BiLSTM and Condition Random Field (CRF) models are used respectively. The models are combined in order to evaluate the performance and progress of railway personnel [37].

B. TEXT-BASED KNOWLEDGE MANAGEMENT FOR RAILWAY DOCUMENTS ANALYSIS

In today's business world, effective knowledge management is essential for all organizations. With the presence of numerous employees scattered across different departments and locations, managing and retaining knowledge becomes a daunting task for large organizations. Additionally, all those documents are in different formats making it even more difficult to access and understand the correct information. This requires the implementation of a centralized knowledge management system (KMS) to ensure that all the organization's best practices and knowledge are readily available to everyone. Research suggests that digital ecosystems can improve human resource management and enhance knowledge, leading to benefits in business performance. It was also found that a knowledge management framework tailored to a firm's size can enhance internal learning processes, and the use of internal and external communication networks, as well as knowledge repositories, can facilitate this process [38].

In recent years, there have been many industry-specific and generalized document analysis tools and KMS. The software module for the automated system of accounting and control of railway automation is implemented for the electronic document management system for technical documentation (EDTD) [39]. Another semantic annotation method was introduced to formalize the technical operation rules of the Ukrainian railway system and information can be ingested from various sources using a modular ontology [40]. All of the industry-specific solutions are based on knowledge graphs or ontology-based techniques that provide very limited information about the context or semantic understanding of the textual knowledge that exists in documents.

OpenAI recently introduced a tool called ChatGPT that is used as a great tool for language generation and text summarization [41]. Google created a chatbot called Google Bard, which utilizes artificial intelligence (AI) to generate human-like conversations with users. This tool utilizes natural language processing and machine learning technologies to understand user inputs, generate appropriate responses, and provide a conversational experience that mimics human interaction [42].

All of these generalized tools are useful for general-purpose daily tasks, but there are some limitations to these systems. They do not have direct access to private or proprietary information stored in enterprise systems. Instead, it only ingests open-source and publicly available information to generate responses to user queries. Another limitation is the availability of these platforms and data protection is key to the railway industry. Specifically in this research, the limitation is the inability to ingest information from sources such as UML (Unified Modeling Language) for EA documents.

III. PURPOSED METHODOLOGY FOR RAILWAY DOCUMENT PROCESSING

The proposed novel document similarity system for the railway safety document processing model is outlined in figure 1. In this research, pre-processing, importing NLP libraries, cleaning of the dataset and tokenization were performed. The documents are tokenized into word and sentence formats. By merging both documents, a database was created that contains each document's name, rules, and clauses. Different pre-trained models were used to embed sentences into vectorized formats for both documents. Each embedded sentence was then compared with the other document to find a similar rule or clause. The similarity between each sentence/clause was measured using cosine similarity metrics. A search feature was also developed that allowed users to search the database for similar rules, clauses, or responsible persons. Searching any rule or clause will display the relevant rule or clause along with its similarity score and document name. The data was visualized using WordCloud, which maps the tokenized words according to their frequency in the railway documents.

IV. DATASETS DETAILS AND TOOLS

This research was conducted on a stand-alone machine with an Intel Xeon W-10885M processor and 64GB RAM. The dataset and other framework details used in this research are explained in the sections below:

A. DATASET FOR RAILWAY DOCUMENT SAFETY

This project is a collaboration between the University of Huddersfield's Institute of Railway Research (IRR) and the RSSB, and the dataset comes from the Enterprise Architecture (EA) of the RSSB, dataset is in the form of CSV file and UML file. One file is called OCD Mapping, a CSV file, and the second one is in UML format called EA-lite. A UML tool for EA called Sparx systems was used for UML analysis as given in figure 2. The main difference between UML and OCD is that UML only contains condensed clauses. On the contrary, there is a detailed version of all the principles and clauses in OCD documents as given in figure 3.

V. NLP TOOLS AND FRAMEWORK

NLP is a process of enabling computers to comprehend and process human language. The main aim is to understand and process understated data to capture key information. Different NLP tools and techniques were utilised to analyse and process this unstructured information. Python is used as a programming language that supports data analysis. Anaconda and Jupyter Notebook was used as a distribution and IDE, respectively.

A. PYTHON

It is a simple yet powerful and interactive programming language. As a platform-independent language, Python runs on all major operating systems. It deals with computing tasks, including data analysis, visualization, and linguistic data efficiently [43]. It has numerous libraries such as Pandas, NumPy, NLTK and others that support NLP and DL.

B. NLTK

In 2001, Natural Language Toolkit (NLTK) was created as an open-source library. This library was developed with four primary principles: uniformity, simplicity, flexibility and adaptability. It provides multiple text corpora, string processing, classification, chunking, parsing, semantic representation, and evaluation metrics [44].

C. TOKENIZATION

It is a process of separating raw text or documents into small chunks such as words or sentences called tokens [45]. It is essential in NLP tasks as it will help to understand the context of the raw corpus. In this work, both documents (EA.csv and OCD.csv) were tokenized into a word and sentence before cleaning the data further.

D. StopWords REMOVAL

Words that contain only minimal information are called stop words [46]. These words are also referred to as noise words

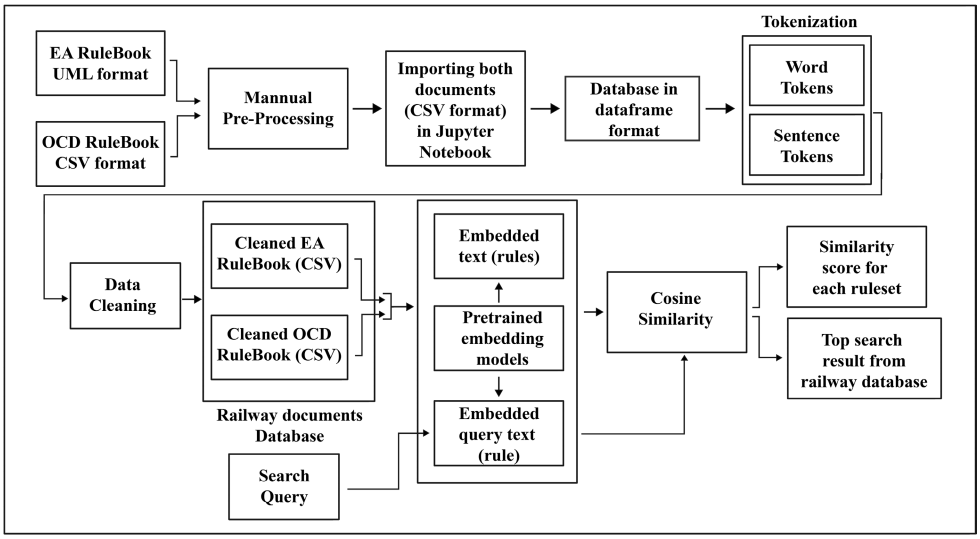

FIGURE 1. Model for railway safety document processing using NLP techniques (such as pre-processing, data cleaning, tokenization, sentence embedding, text similarity measurement using Cosine similarity metrics.)

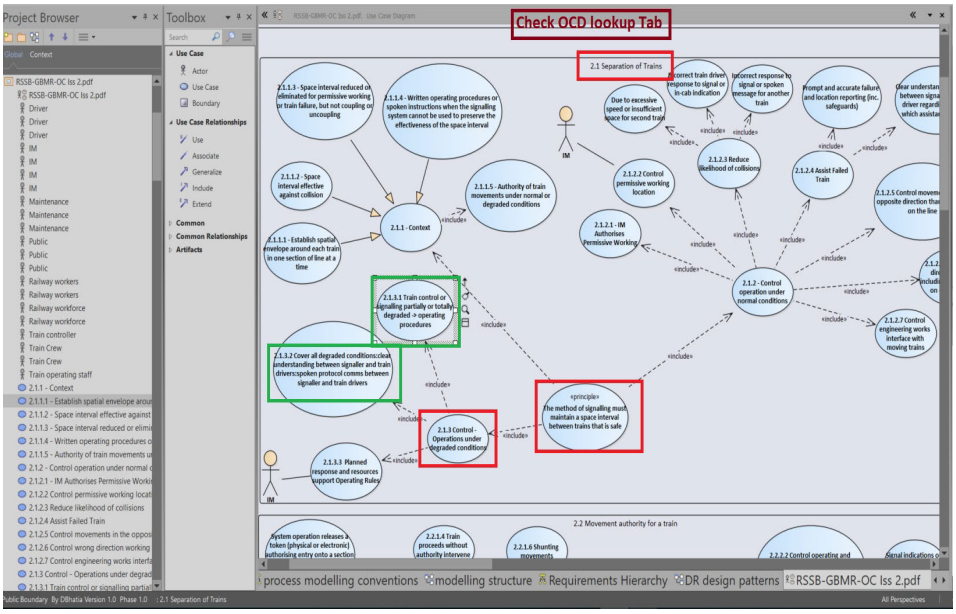

FIGURE 2. Visualization of Enterprise Architecture (EA) document using SPARX SYSTEMS software before pre-processing.

since they don't add any meaning to the sentences. By removing these words from the corpus, the processing will remain the same for the desired task. Dimension of the corpus and noise from the vocabulary can be reduced by removing these words. This will increase the density of the corpus and improve the speed and efficiency of the analysis [48].

E. WordCloud

A WordCloud is a visual representation of all the words that appear most frequently in a corpus or a set of documents [47]. It is helpful to understand the corpus as different words are pictures in different sizes and colours based on their frequency in the documents. WordCloud is used in this study to show the frequency of the most used words in both documents (EA and OCD) as shown in figure 7.

VI. SENTENCE EMBEDDING TECHNIQUES

There are different sentence embedding techniques that play an important role in any NLP task. The year 2018 is considered a critical point for NLP as Google introduced a new language model known as Bidirectional Encoder Representation from Transformers (BERT). DistilBERT Base is a smaller general-purpose language representation model technique. It can be used for a variety of tasks as other transformer

	2.1	**Separation of trains**
1		**Principle: The method of signalling must maintain a space interval between trains that is safe.**
	2.1.2	**Controls - Operation under normal conditions**
	2.1.2.1	Train control and signalling systems must be supported by operating rules, as well as procedures specific to each system, to enable operators to maintain a safe space interval between trains within their area of control when the system is operating under normal conditions.
	2.1.2.2	Permissive working must be limited to locations where it is authorised by the infrastructure manager. In the case of trains conveying passengers, permissive working must be limited to stations, for the purpose of joining trains or platform sharing.
	2.1.2.3	Operating rules for permissive working must be designed to reduce the likelihood of: a) collisions due to excessive speed during movements or insufficient space for the second train b) incorrect response by a train driver to a signal or in-cab indication c) incorrect response by a train driver to a signal or spoken message intended for another train.
	2.1.2.4	Operating rules for assisting a failed train (where permissive working is not available), must include: a) a requirement for prompt and accurate reporting of the failure and its location b) a requirement for establishing a clear understanding between the signaller and train driver of the failed train about the direction from which assistance will come c) safeguards against possible errors in locating the failed train d) protocols for spoken communications between signaller and train drivers.
	2.1.2.5	Operating rules must be applied to control movements in the opposite direction to that for which the line is signalled, to control the risks of collision and derailment from: a) unsignalled wrong-direction movements including single line working b) movements between one line and another at each end of the section being used for the wrong direction movement c) communication to train drivers of insufficient or incorrect information about movements to be made and conditions to be applied.
	2.1.2.6	Wrong direction working procedures must take account of people working on or near the line being used, including those employed to display hand signals controlling train movements. The management of risks to people from moving trains is covered in section 2.8 of this operational concept document.
	2.1.2.7	The interface between engineering works and moving trains is covered in section 2.3 of this operational concept document.

Tranches | Modules | Principles | OCD lookup | OCD | Comments | Index | M1 | M2 | M3 | P1 | P2 | PoSA | S4 | S5 | S7 | SP | SS1 | SS2 | ...

FIGURE 3. Visualization of Operational Concept Document (OCD) before pre-processing.

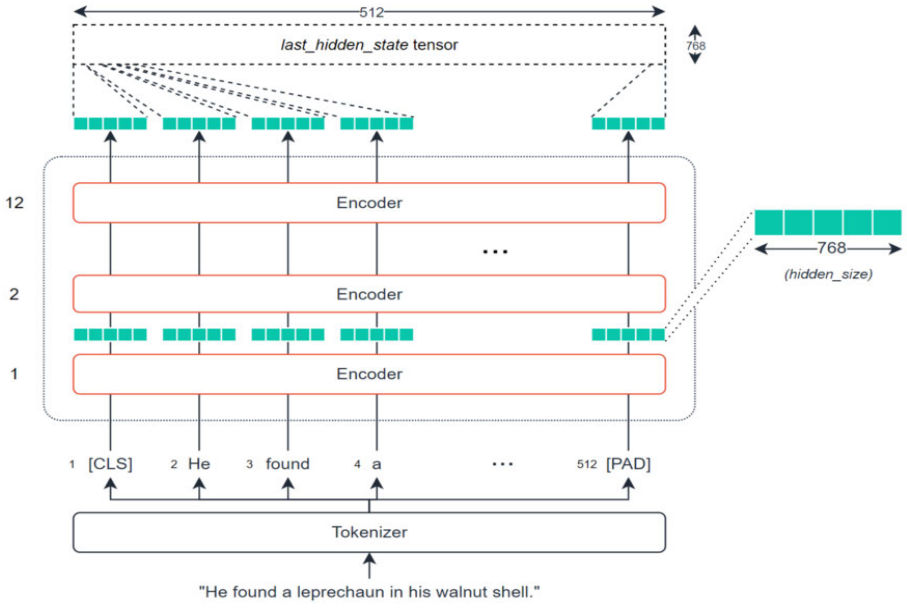

FIGURE 4. BERT base network architecture [49].

models. Its architecture is similar to BERT, but token embedding and pools are removed with several layers less by a factor of two. All other vital operations such as linear layer and layer normalisation are improved using model linear algebra frameworks [55]. Electra Small is another pre-trained transformer model introduced by Google. It maps the sentences into dense vector space. Embedded vectors can be used in a variety of NLP tasks, such as clustering and abstracting [57].

A. BERT - BIDIRECTIONAL ENCODER REPRESENTATION FROM TRANSFORMERS

BERT is one of the most advanced ML frameworks. Its design was based on previous research including Semi-supervised sequence learning [50], ELmO [51] and ULMFit [52]. Unlike the earlier models, BERT is multi-layer bidirectional transformer layers that attenuate in both directions [53]. There are two variants of the BERT model released by Google: BERT_{base} and BERT_{large}. In the BERT_{base} model, there are twelve transformer layers, while in the BERT_{large} model there are twenty-four layers. These models also have more extensive feed-forward networks of 768 and 1024 hidden units and attenuation heads of twelve and sixteen, respectively.

BERT employs two unsupervised strategies named Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). In MLM, a certain percentage of the words in each sentence are replaced with masked tokens. Based on the context of the surrounding words, the model will predict those masked tokens. A conventional language model can only be trained in one direction, resulting in trivial predictions of the target words. MLM model is trained on bidirectional representation resulting in a deep language context. NSP is based on capturing the relationship between two sentences.

To understand the relationship BERT uses pairs of sentences for training. For training, a dataset is divided into two parts. In the first part, sentences A and B for each training example are the subsequent sentences and are labelled as 'IsNext'. For the remaining training dataset, sentence A is paired with random sentence B with the 'NotNext' label. This pre-training is very beneficial in Question Answering and Natural language Interface [55].

In this research, Electra Small, DistillBERT Base and BERT Base pre-trained models were used to embed the sentences (rules, principles, clauses or responsible actor) into a vectorized format. These sentences are further compared using cosine similarity metrics to compute the similarity.

VII. SIMILARITY METRICS MODEL

A. COSINE SIMILARITY

A cosine similarity metric is the measurement of the angle between two or more vectors projected in multidimensional space. It is commonly used in document classification, information retrieval and document similarity measurement. Documents or sentences can be embedded into a vectorized format using embedding techniques (such as the word2vec or BERT) and the angle between the embedded sentences can be measured [58]. The angle between vectorized embeddings indicates the similarity between sentences of the documents. A smaller angle between vectorized embeddings produces a large cosine angle indicating a higher cosine similarity, as illustrated in the figure 5. Mathematically it can be described as the dot product of vectorized embedded sentences and the product magnitude of each vectorized sentence and is calculated using equation 1.

$$\cos(Sent1, Sent2) = \frac{\sum_{i=1}^{n} Sent1_i Sent2_i}{\sqrt{\sum_{i=1}^{n} (Sent1_i)^2}\sqrt{\sum_{i=1}^{n} (Sent2_i)^2}} \tag{1}$$

In equation 1, $Sent1$ and $Sent2$ represents two vectorized sentences from the railway safety documents. $Sent1.Sent2$ are the scalar product of two vectorized sentences, whereas $\sqrt{\sum_{i=1}^{n} (Sent1_i)^2}$ and $\sqrt{\sum_{i=1}^{n} (Sent2_i)^2}$ is the product of the euclidean norms of the vectorized sentence.

VIII. EXPERIMENTS AND RESULTS

This research aims to develop a framework for a document safety system for the railway. RSSB has provided railway safety documents (EA and OCD). The semantic similarity between the two documents was analyzed. A search function was also developed that allows users to search through multiple documents, check the rule's existence, and amend or create new documents accordingly. Various NLP tools and libraries were used for document processing. A detailed process of documents analysis is explained below:

1) **Mannual Pre-processing:** Both railway safety documents provided by RSSB (EA and OCD) were in different formats and manually converted into .CSV format. EA document was in UML format and contains free text. This was added manually to the EA document in CSV format.
2) **Importing the documents:** Both documents were imported into Jupyter Notebook using Python. OCD Rulebook document contains rules and principles that are comprehensive and elaborate. On the other hand, the EA Rulebook document consists of was brief and concise.
3) **Database:** Both documents were combined into a single database using the Pandas data frame. This data frame consists of the document name and its rules and clauses.
4) **Document Tokenization:** The NLTK library was used to tokenize documents into words and sentences.
5) **Cleaning the documents:** All clauses that do not start with sequence numbers were removed. To reduce dimensionality and noise, stop words such as ''a'', ''is'', and ''the'' were also removed.
6) **Embedding documents:** For embedding document sentences into the vectorized format, three pre-trained models Electra Small, DistilBERT Base, and BERT Base were used.
7) **Document similarity:** The cosine similarity metric is used to map the semantic similarity between each rule of the EA with OCD documents to find the relevant match.
8) **Search method:** search feature was developed that let the user search using a search string from both the documents and display the semantically similar match rule, clause or principle along with similarity score and document name.

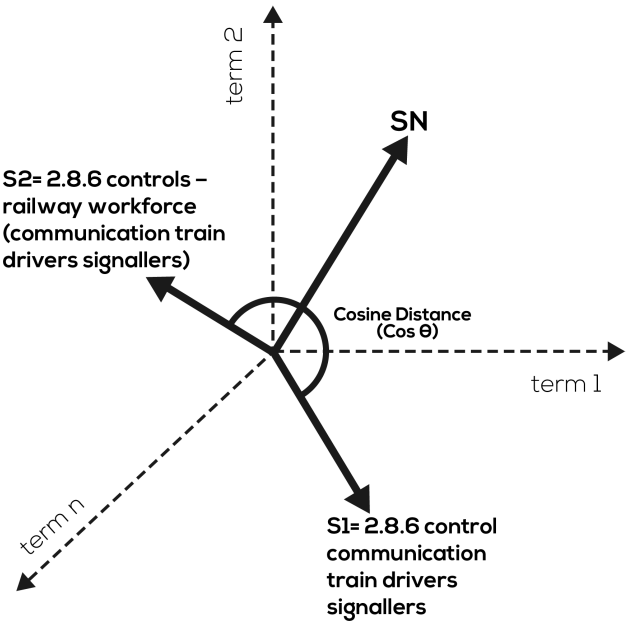

FIGURE 5. Cosine similarity measurement between two embedded sentences of railway safety documents.

```
#printing the embedding size and shape
sentence_embeddings1 = model.encode(source_rules)
print(source_rules[0])
print(sentence_embeddings1.shape)
print(sentence_embeddings1[0])
```

```
2.1.3 controls – operations un- der degraded conditions
(6, 768)
[-0.04111726  0.85672855  1.1065563   0.11842521  0.29113722 -0.5358921
  1.5732343  -0.39404038  0.21237323  0.18504868 -1.2871532   0.6570389
 -0.3492697  -0.27618137  0.10996132  1.0760059  -0.7678367  -0.89649093
  0.55660385 -0.63322145 -0.2055838  -0.17879556 -0.62644863 -0.3646943
 -0.02013143 -0.07552358  0.08249988  0.08896728 -1.545986   -0.04053611
  0.30826226  0.03533401  0.07240922 -0.34463936  0.6073419   1.5145692
  0.32344502  0.12372848 -0.07149942 -0.12006197  1.7055484   0.0644784
  0.59089667 -0.03787512 -0.32094383 -0.08943181  0.646159    0.10397064
 -0.02832493 -0.26395994 -0.51379883  0.7722048   0.38389942  0.5776116
 -0.14818367 -0.61892486  0.4025852  -0.67636055 -0.20911439 -0.08390769
  0.2766656  -0.5380783   0.3041315   0.7890702  -0.96928376  0.08387748
 -0.5693102  -0.10262232 -0.6372029  -0.36283058  0.5636814  -0.47335
 -0.5358241   0.21781412 -0.28557464 -0.6021847  -0.11898013  0.16580582
 -0.3088822   0.7976688   0.4710376   0.69001347  1.2996045  -0.4503242
 -0.32533795 -0.0051183   0.9867227   0.5906702  -1.3404758   0.6604217
  0.79081666  0.9834364   1.0955485  -0.4487591  -0.11101138  0.1078865
  0.0302773   0.05716512  1.0621036  -0.04783775 -0.47829896 -0.09535808
```

FIGURE 6. Visualization of EA '2.1.3 rule' in 6 × 768 vectorized cross-sectional format.

9) **Visualization:** term frequency of the most frequent words used in the documents is visualized using the WordCloud library. This provides railway employees and key personnel with valuable insights and key trends in the documents.

A. SENTENCE EMBEDDING VISUALIZATION

In order to embed the sentences in the documents, different pre-trained models are used. The EA rule 2.1.3 is visualized to show the embedding shape and cross-section of the vectorized format. BERT model embeds the sentence into a 6 × 768 dimension as shown in figure 6.

B. WordCloud: TERM FREQUENCY for DOCUMENTS

In this research, WordCloud is used to visualize the most frequent words in the documents [56]. A cloud of words is created to show the highlighted words and phrases. Based on frequency and relevance, words are shown in a bigger size in WordCloud. Train, driver and signaller are the most used words in both documents, as shown in figure 7.

Its development was specifically intended to enhance the comprehension of commonly used words in safety documents among railway personnel, enterprise architects, and other stakeholders. Furthermore, the real-time dashboard associated with this tool can also be used to provide valuable insights and trend analysis.

C. SEMANTIC SIMILARITY FOR RAILWAY DOCUMENTS

The semantic similarity between railway safety documents is measured utilizing the cosine similarity metrics. Semantic similarity between the embedded sentences (or rules) ranges from 0 to 1, where a value of 1 signifies the highest degree of similarity between two clauses, while 0 denotes the absence of similarity. The threshold to map the semantic similarity between the rules of two documents is set to 50%. The threshold value can be adjusted according to requirements. Each ruleset from the EA document is compared with the OCD document to check the relevant rule existence. Positive and negative signs in table 1 show the correct or incorrect mapping of the two documents respectively.

In this research, three pre-trained embedding models named Electra Small, DistilBERT Base and BERT Base were utilized to measure the semantic similarity of railway safety documents. It also allows us to compare the different embedding model's performance for railway safety critical documents analysis.

By using Electra Small as an embedding model, EA document rule 2.8.5 was mapped accurately to the OCD document rule resulting in a positive similarity score of 0.95. However, for the remaining EA document rules such as 2.1.3.2, 2.4.1.1, 2.5.2.2, and 2.2.1.6, the Electra Small embedding model was unable to correctly map it to the OCD document rules. This shows that Electra Small embedding model is only effective for short and brief rules. However, when the rules are comprehensive and detailed it was not able to find the correct match.

DistilBERT Base show promising results as compared to Electra Small embedding model. When comparing EA rule 2.1.3.2, it correctly matches the OCD rule with a positive relevance score of 0.68 whereas Electra Small was unable to map the correct match. For EA rule 2.4.1.1, DistilBERT Base and BERT Base show positive similarity scores of 0.81 and 0.84 respectively. When mapping EA rule 2.5.2.2 to OCD document rules, BERT Base shows a maximum relevance score of 0.89 with positive similarity. Similarity for EA rule "2.2.1.6 Shunting movement" accurately maps to OCD rule 2.2.1.6 with a highest relevance score of 0.63.

BERT Base outperforms the other embedding models for railway document processing. It outperforms other embedding models in terms of relevance score and correct mapping of rules across the documents.

D. SEARCH METHOD FOR RAILWAY DOCUMENTS

During this research, a semantic search feature was developed that allows users to search through the documents. All the rules from EA and OCD documents are combined in one database in a data frame format using pandas. A user can then search for any rule using a search string. Based on the semantic context, the system will provide the most relevant matching rules or principles. Additionally, it will display the similarity score and the document name. As a result, enterprise architects and other key railway personnel will be able to avoid duplication of rules across existing or new documentation.

All sentences from the data frame and search string are embedded using the BERT Base model. To find the relevant rule or clause, cosine similarity metrics are applied to the embedded search string and all other embedding sentences. The search system is set to show the top 5 relevant rules with the best similarity score in descending order along with the document name. When the search string "principle: the method of signalling" is searched to find the relevant match in the database. The top result is "Principle: the method of

FIGURE 7. Visual term frequency representation for railway safety documents using WordCloud.

TABLE 1. Comparison based on different embedding techniques for railway document processing.

EA Document	OCD Document	Electra Small	DistilBERT Base	BERT Base
2.1.3.2 Cover all degraded conditions: clear understanding between signaller and train drivers: spoken protocol comms between signaller and train drivers	2.1.3.2 the rules must provide for: a) the full range of foreseeable failure conditions, from simple to complex, and their consequences for the functionality of the system as a whole and its ability to maintain a safe interval between trains b) establishing a clear understanding between signallers and train drivers about movements to be made and precautions to be taken during the movement to reduce the likelihood of collision c) protocols for spoken communications between signaller and train drivers and any intermediaries involved in the passing of information.	0.92 (-ive)	0.68 (+ive)	0.83 (+ive)
2.4.1.1 Controls prevent threat to train movements presented by an obstruction or unsafe condition	2.4.1.1 the starting point of this principle is that a section of line is known or suspected to be unsafe for trains to pass over it. controls to ensure the safety of the infrastructure for train movements will have been overcome or circumvented; the controls described here are designed to prevent the threat to train movements presented by an obstruction or unsafe condition being realised.	0.94 (-ive)	0.81 (+ive)	0.84 (+ive)
2.5.2.2 Trains without automatic brakes or through pipes cannot move on the mainline network without additional safeguards applied	2.5.2.2 vehicles without operative automatic brakes or through pipes (including locomotives, on-track machines and multiple units being hauled unpowered in a freight train) must not be moved on the mainline network unless additional safeguards are applied, which must be published in operating rules and operating instructions.	0.88 (-ive)	0.82 (+ive)	0.89 (+ive)
2.5.3.5 Mitigate risk of an incident for dangerous goods	2.5.3.5 to mitigate the risks in the event of an incident involving dangerous goods, operating rules for those responsible for freight train preparation and operation, and traffic control, must define requirements for: a) separating dangerous goods from other traffic b) reporting and managing dangerous goods incidents.	0.84 (-ive)	0.57 (+ive)	0.69 (+ive)
2.8.5 Control rail vehicle examination or repair	2.8.5 controls - railway workforce (examination or repair of a train or rail vehicle)	0.95 (+ive)	0.73 (+ive)	0.78 (+ive)
2.2.1.6 Shunting movements	2.2.1.6 for some shunting movements, the authority and its extent may be conveyed entirely by spoken instructions or hand signals, which may also be used to provide continuing confirmation of the authority.	0.73 (-ive)	0.50 (+ive)	0.63 (+ive)

signalling must maintain a space interval between trains that is safe'' with a similarity score of 0.94. This rule is from the OCD rulebook as shown in figure 8. Similarly when ''2.6.2.4 Operating rules must include:'' was searched as a

	Matching Sentence	Similarity Score	Document name
1	principle: the method of signalling must maintain a space interval between trains that is safe.	0.940703	Ocd Rulebook.Csv
52	principle: trains proceeding over any portion of line must not be obstructed in a way that threatens their safety.	0.903734	Ocd Rulebook.Csv
494	the method of signalling must maintain a space interval between trains that is safe	0.894346	Ea Rulebook.Csv
282	principle: the workforce must be protected from the particular hazards associated with electrified railways.	0.886092	Ocd Rulebook.Csv
23	" principle: before a train is allowed to start or continue moving, it must have an authority to move that clearly indicates the limit of that authority."	0.880589	Ocd Rulebook.Csv

FIGURE 8. Search result for a "principle clause" from railway documents with similarity score and document name.

	Matching Sentence	Similarity Score	Document name
134	2.6.2.4 operating rules includes vehicle and infrastructure constraints and procedures	0.835337	Ea Rulebook.Csv
308	2.3.1.4 operating rules are used:	0.811807	Ocd Rulebook.Csv
533	2.9.1.4 the additional controls involve:	0.799034	Ocd Rulebook.Csv
206	operating rules and procedures	0.773961	Ea Rulebook.Csv
60	2.3.1.4 level crossing operating rules	0.769343	Ea Rulebook.Csv

FIGURE 9. Search result for a search string "2.6.2.4 Operating rules must include:" from railway documents with similarity score and document name.

string, EA rulebook 2.6.2.4 was shown as the most relevant rule with a similarity score of 0.83 as shown in figure 9.

IX. CONCLUSION

This research demonstrated a novel document processing method for railway safety documents. It investigates NLP's effectiveness in verifying safety rules within the EA model. It also focuses on the possibility of automatically updating EA models using the RSSB Rule Book and ensures all relevant information is included and up-to-date. Secondly, a semantic search method was developed during this research so that railway personnel will be able to confirm the presence of rules or principles in both railway documents based on the semantic context. It will also provide a consistent check on the railway documents.

During this research two (EA and OCD) documents were considered where EA contained brief rules and OCD contained detailed rules and principles. Both documents were manually pre-processed and imported into Jupyter Notebook and databases were created using the pandas dataframe. Rules in the documents were tokenised, cleaned, and embedded into a vectorized format. Three pre-trained embeddings and the cosine similarity metrics were utilized in order to perform document processing. The semantic similarity between document sentences ranges from 0 to 1.

Electra Small pre-trained model results were unsatisfactory as the embeddings were less dense with 256 dimensions for a sentence. It was only able to map short and brief rules across the documents whereas when rules were long and detailed it was unable to correctly map document rules as illustrated in the table 1. Dense vector space embeddings were provided by DistilBERT Base and BERT Base. Both pre-trained models demonstrate more accurate results for document similarity. For EA rule 2.5.2.2, both DistilBERT Base and BERT Base show positive similarity with a relevance score of 0.82 and 0.89. Similarly for EA 2.2.1.6, BERT Base outperform the DistilBERT Base embedding model with a maximum relevance score of 0.63. This shows that the BERT Base embedding model outperforms DistilBERT Base with more accurate results in terms of accuracy. A search method using the BERT Base model was devised that displays relevant rules or principles from railway documents along with the document name and similarity score. A Worcloud was also developed that provides insights and key trends in railway documents.

To conclude, this research is conducted for railway safety critical document processing and the system developed in this research allows railway employees and stakeholders to devise information about railway documents in a useful manner. With minor tweaks in the data ingestion and cleaning process, the document processing system developed in this research can be used as a text-based KMS system for various industries.

X. FUTURE WORK

In the future, topic modelling can be explored to process railway documents. Based on the words and phrases in the documents, topic modelling will group documents into clusters. It is possible to cluster documents allowing users to read through specific topics rather than all available documents. A web-based dashboard system will be developed that will allow all the railway enterprise architects and other key workers to efficiently search for the existence of any rules in the database.

ACKNOWLEDGMENT

The authors would like to thank the Institute of Railway Research, University of Huddersfield, and Rail Safety and Standards Board (RSSB) for providing this dataset and feedback that allows us to improve the systems.

REFERENCES

[1] S. Bernard, *An Introduction to Enterprise Architecture*. Bloomington, IN, USA: AuthorHouse, 2012.

[2] F. Goethals, "An overview of enterprise architecture framework deliverables," *SSRN Electron. J.*, vol. 870207, Dec. 2005.

[3] K. W. Burrage, "Railway safety standards," in *Proc. Int. Conf. Electr. Railways United Eur.*, 1995, pp. 153–157.

[4] K. Schuitemaker and G. M. Bonnema, "Modelling integral risk assessment (MOIRA): Experiments on the Dutch railway departure process," in *Proc. 14th Annu. Conf. Syst. Syst. Eng. (SoSE)*, May 2019, pp. 272–277.

[5] A. Canning, "Risk and safety decision making," *IET Prof. Develop. Course Electr. Traction Syst.*, pp. 185–190, Nov. 2010.

[6] H. Parkinson and G. Bamford, "A journey into railway digitisation," in *Proc. Stephenson Conf., Res. Railways*, 2017, pp. 333–340.

[7] H. Alawad, S. Kaewunruen, and M. An, "A deep learning approach towards railway safety risk assessment," *IEEE Access*, vol. 8, pp. 102811–102832, 2020.

[8] R. Mihalcea, C. Corley, and C. Strapparava, "Others Corpus-based and knowledge-based measures of text semantic similarity," in *Proc. AAAI*, vol. 6, 2006, pp. 775–780.

[9] S. A. Al-Ghamdi, J. Khabti, and H. S. Al-Khalifa, "Exploring NLP web APIs for building Arabic systems," in *Proc. 12th Int. Conf. Digit. Inf. Manage. (ICDIM)*, Sep. 2017, pp. 175–178.

[10] T. Ganegedara, *Natural Language Processing With TensorFlow: Teach Language to Machines Using Python's Deep Learning Library*. Birmingham, U.K.: Packt Publishing Ltd, 2018, pp. 1–2.

[11] O. Levy and Y. Goldberg, "Dependency-based word embeddings," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics (Short Papers)*, vol. 2, 2014, pp. 302–308.

[12] X. Zhang and Y. LeCun, "Which encoding is the best for text classification in Chinese, english, Japanese and Korean?" 2017, *arXiv:1708.02657*.

[13] I. Arroyo-Fernández, C.-F. Méndez-Cruz, G. Sierra, J.-M. Torres-Moreno, and G. Sidorov, "Unsupervised sentence representations as word information series: Revisiting TF–IDF," *Comput. Speech Lang.*, vol. 56, pp. 107–129, Jul. 2019.

[14] Y. Goldberg and O. Levy, "word2vec explained: Deriving Mikolov et al.'s negative-sampling word-embedding method," 2014, *arXiv:1402.3722*.

[15] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, *arXiv:1301.3781*.

[16] H. Yao, H. Liu, and P. Zhang, "A novel sentence similarity model with word embedding based on convolutional neural network," *Concurrency Comput., Pract. Exper.*, vol. 30, no. 23, Dec. 2018, Art. no. e4415.

[17] J. Xiao and Z. Zhou, "Research progress of RNN language model," in *Proc. IEEE Int. Conf. Artif. Intell. Comput. Appl. (ICAICA)*, Jun. 2020, pp. 1285–1288.

[18] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, pp. 1735–1780, Jan. 1997.

[19] J. Müeller and A. Thyagarajan, "Siamese recurrent architectures for learning sentence similarity," in *Proc. 13th AAAI Conf. Artif. Intell.*, 2016, pp. 1–7.

[20] Y. Huang, Y. Jiang, T. Hasan, Q. Jiang, and C. Li, "A topic BiLSTM model for sentiment classification," in *Proc. 2nd Int. Conf. Innov. Artif. Intell.*, Mar. 2018, pp. 143–147.

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5999–6009.

[22] *IEEE Recommended Practice for Functional Testing of a Communications-Based Train Control (CBTC) System*, Standard 1474.4-2011, Institute of Electrical and Electronics Engineers, 2011, pp. 1–46.

[23] *Railway Applications-Specification and Demonstration of Reliability*, document IEC 62278, International Electrotechnical Commission, Availability, Maintainability and Safety (RAMS), 2020. [Online]. Available: Https://websTOre.Iec.Ch/publiCatio

[24] *Railway Applications—The Specification and Demonstration of Reliability, Availability, Maintainability and Safety (RAMS)*, document CENELEC, EN 50126, Brussels, Belgium, 2007.

[25] *Railway Applications Communication, Signalling and Processing Systems-Software for Railway Control and Protection Systems*, document CENELEC, EN 50128, Brussels, Belgium, 2020.

[26] *Railway Applications Communication, Signalling and Processing Systems Safety Related Electronic Systems for Signalling*, document CENELEC, EN 50129, Brussels, Belgium, 2018.

[27] E. Selig, G. Cardillo, E. Stephens, and A. Smith, "Analyzing and forecasting railway data using linear data analysis," *WIT Trans. Built Environ.*, vol. 103, pp. 25–34, Sep. 2008.

[28] I. Gokasar and K. Simsek, *Using Big Data for Analysis and Improvement of Public Transportation Systems in Istanbul*. Athens, Greece: Academy of Science, 2014.

[29] A. Thaduri, D. Galar, and U. Kumar, "Railway assets: A potential domain for big data analytics," *Proc. Comput. Sci.*, vol. 53, pp. 457–467, Jan. 2015.

[30] M. Faizrahnemoon, A. Schlote, L. Maggi, E. Crisostomi, and R. Shorten, "A big-data model for multi-modal public transportation with application to macroscopic control and optimisation," *Int. J. Control*, vol. 88, no. 11, pp. 2354–2368, Nov. 2015.

[31] D. Lopez. *Your Guide to Natural Language Processing (NLP)*. Medium. Accessed: Jun. 2019. [Online]. Available: https://towardsdatascience.com/your-guide-to-natural-language-processing-nlp-48ea2511f6e1

[32] P. Hughes, R. Robinson, M. Figueres-Esteban, and C. van Gulijk, "Extracting safety information from multi-lingual accident reports using an ontology-based approach," *Saf. Sci.*, vol. 118, pp. 288–297, Oct. 2019.

[33] J. Fei, G. Ni, X. Wang, and Y. Jie, "Cluster analysis on railway infrastructure standards in China and its application to railway efficiency improvement," *IOP Conf. Ser., Earth Environ. Sci.*, vol. 467, no. 1, Mar. 2020, Art. no. 012191, doi: 10.1088/1755-1315/467/1/012191.

[34] M. Figueres-Esteban, P. Hughes, and C. van Gulijk, "Visual analytics for text-based railway incident reports," *Saf. Sci.*, vol. 89, pp. 72–76, Nov. 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0925753516300832

[35] K. Syeda, S. Shirazi, S. Naqvi, and H. Parkinson, and G. Bamford, "Big data and natural language processing for analysing railway safety: Analysis of railway incident reports," in *Human Performance Technology: Concepts, Methodologies, Tools, and Applications*. Hershey, PA, USA: IGI Global, 2019, pp. 781–809.

[36] S. Yavuz and A Yazıcı, "Named entity recognition in Turkish with Bayesian learning and hybrid approaches," in *Proc. Inf. Sci. Syst., 28th Int. Symp. Comput. Inf. Sci.*, 2013, pp. 129–138.

[37] F. Gao, S. Wang, X. Li, H. Cao, and X. Cai, "Research on text mining of railway safety supervisors performance based on BiLSTM and CRF," *J. Phys., Conf. Ser.*, vol. 1213, no. 5, Jun. 2019, Art. no. 052016.

[38] A. de Bem Machado, S. Secinaro, D. Calandra, and F. Lanzalonga, "Knowledge management and digital transformation for industry 4.0: A structured literature review," *Knowl. Manage. Res. Pract.*, vol. 20, no. 2, pp. 320–338, Mar. 2022.

[39] A. Ogli, "Software for electronic document management system of technical documentation on railway automation and telemechanics," *JournalNX*, vol. 7, no. 1, pp. 204–209, 2021.

[40] V. Shynkarenko and L. Zhuchyi, "Semantic checking of different type information sources about permitted speeds in railway transport," in *Proc. CEUR Workshop*, vol. 3171, 2022, pp. 711–723.

[41] M. Abdullah, A. Madain, and Y. Jararweh, "ChatGPT: Fundamentals, applications and social impacts," in *Proc. 9th Int. Conf. Social Netw. Anal., Manage. Secur. (SNAMS)*, Nov. 2022, pp. 1–8.

[42] M. Rahaman, M. Ahsan, N. Anjum, M. Rahman, and M. Rahman, "The AI race is on! Google's bard and OpenAI's ChatGPT head to head: An opinion article," in *The AI Race is on*, M. Rahman and M. Nafizur, Eds. SSRN (Social Science Research Network), 2023.

[43] J. Vanderplas, *Python Data Science Handbook: Tools and Techniques for Developers*. Sebastopol, CA, USA: O'Reilly, 2016.

[44] S. Bird, E. Klein, and E. Loper, *Natural Language Processing With Python: Analyzing Text With the Natural Language Toolkit*. Sebastopol, CA, USA: O'Reilly Media, Inc., 2009.

[45] Y. A. Solangi, Z. A. Solangi, S. Aarain, A. Abro, G. A. Mallah, and A. Shah, "Review on natural language processing (NLP) and its toolkits for opinion mining and sentiment analysis," in *Proc. IEEE 5th Int. Conf. Eng. Technol. Appl. Sci. (ICETAS)*, Nov. 2018, pp. 1–4.

[46] V. Jha, N. Manjunath, P. D. Shenoy, and K. R. Venugopal, "HSRA: Hindi stopword removal algorithm," in *Proc. Int. Conf. Microelectron., Comput. Commun. (MicroCom)*, Jan. 2016, pp. 1–5.
[47] M. F. A. Bashri and R. Kusumaningrum, "Sentiment analysis using latent Dirichlet allocation and topic polarity wordcloud visualization," in *Proc. 5th Int. Conf. Inf. Commun. Technol. (ICoIC7)*, May 2017, pp. 1–5.
[48] J. Kaur and P. Buttar, "A systematic review on stopword removal algorithms," *Int. J. Future Revolution Comput. Sci. Commun. Eng.*, vol. 4, pp. 207–210, Apr. 2018.
[49] J. Briggs. (Sep. 2021). *Bert for Measuring Text Similarity*. Medium. [Online]. Available: https://towardsdatascience.com/bert-for-measuring-text-similarity-eec91c6bf9e1
[50] A. M. Dai and Q. V. Le, "Semi-supervised sequence learning," 2015, *arXiv:1511.01432*.
[51] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," 2018, *arXiv:1802.05365*.
[52] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," 2018, *arXiv:1801.06146*.
[53] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2019, *arXiv:1810.04805*.
[54] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," 2019, *arXiv:1910.01108*.
[55] Y. Jiang, B. Sharma, M. Madhavi, and H. Li, "Knowledge distillation from BERT transformer to speech transformer for intent classification," 2021, *arXiv:2108.02598*.
[56] F. Tolner, G. Eigner, B. Barta, and M. Takacs, "Investigation of high-growth firms in the SME sector via the perspective of owners and CEOs using wordclouds," in *Proc. IEEE 15th Int. Symp. Appl. Comput. Intell. Informat. (SACI)*, May 2021, pp. 375–380.
[57] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, "ELECTRA: Pre-training text encoders as discriminators rather than generators," 2020, *arXiv:2003.10555*.
[58] F. Rahutomo, T. Kitasuka, and M. Aritsugi, "Semantic cosine similarity," in *Proc. 7th Int. Student Conf. Adv. Sci. Technol. (ICAST)*, vol. 4, 2012, p. 1.

ABDUL WAHAB QURASHI is currently pursuing the Ph.D. degree in natural language processing and high-performance computing from the University of Huddersfield. He is a Senior Software Engineer with the Advanced Manufacturing Research Centre (AMRC), The University of Sheffield, U.K. His research interests include natural language processing (NLP), data analysis, high-performance computing (HPC), and networking.

ZOHAIB A. FARHAT received the B.Sc. degree in electrical engineering from Air University, Pakistan, and the M.Sc. degree in embedded systems and the Ph.D. degree in visible light communication and intensity modulation (PPM) from the University of Huddersfield, U.K. He is currently a Senior Embedded Systems Engineer with the Advanced Manufacturing Research Centre (AMRC), The University of Sheffield, U.K. He is also involved in different projects, including smart manufacturing, designing customized sensors for 5G communication, Industry 4.0, and knowledge management systems.

VIOLETA HOLMES received the bachelor's degree in electronics and physics from the ETF Electronic and Technical Physics Faculty, University of Belgrade, Yugoslavia, the M.Sc. degree in control engineering from the University of Bradford, and the Ph.D. degree in the application of artificial intelligence in distributed computer control from the University of Huddersfield. In 2007, she joined the School of Computing and Engineering, University of Huddersfield, as a Senior Lecturer in DSP and embedded systems. She lectured in parallel computer architectures, computer clusters, cloud and grid, embedded systems, digital signal processing, and virtual instrumentation. She was the Course Leader for the Internet of Things (M.Sc.), the Embedded Systems Engineering (M.Sc.), and the Electronic and Communication Engineering (M.Sc.). She was also a Systems Engineer with Smederevo Steel Mill, Yugoslavia, gaining industrial experience in computer control of hot and cold steel rolling mills, and a Reader in high-performance computing with the University of Huddersfield, for more than 25 years of teaching and research experience in computing and engineering. She led the High-Performance Computing (HPC) Research Group, University of Huddersfield. She was an ARCHER Champion and a Deep Learning Institute Certified Instructor. Her research interests include HPC systems infrastructure, computer clusters, grids, cloud computing, intelligent agents, big data, the Internet of Things, and embedded systems. She was a member of the Institute of Engineering and Technology (IET) and the British Computer Society (BCS). She was awarded the status of Chartered Engineer and the Fellowship of Higher Education Academy.

ANJU P. JOHNSON received the Ph.D. degree in cybersecurity from the Department of Computer Science and Engineering, Indian Institute of Technology (IIT) Kharagpur. She is currently a Senior Lecturer in electronic engineering and embedded systems with the University of Huddersfield, with more than ten years of teaching and research experience in computing and engineering in the U.K. and abroad. She is also an External Examiner with the School of Computing, Engineering and Intelligent Systems, Derry Campus, Ulster University, U.K., for the B.Eng. degree (Hons.) in electrical and electronic engineering and the B.Eng. degree (Hons.) in computer engineering programs. Previously, she worked as a Postdoctoral Research Associate with the Department of Electronic Engineering, University of York, U.K., from 2016 to 2018. Before joining the University of York, she was a Senior Project Officer in the Sponsored Research and Industrial Consultancy with the Indian Institute of Technology (IIT) Kharagpur, from 2012 to 2016, and a Lecturer with the Department of Electronics and Communication Engineering, National Institute of Technology (NIT) Calicut, India, in 2012. She has close to 30 publications in reputed international journals and conferences. In addition, she has rendered her service as a reviewer and program committee member for multiple workshops, conferences, and journals. Her research interests include the Internet of Things, hardware security, neuromorphic computing, and field-programmable gate array prototyping. She was one of the recipients of the President of India (Rashtrapati) Award, in 2006, for her services to the country.