IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 5 March 2024, accepted 11 March 2024, date of publication 18 March 2024, date of current version 22 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3377232

RESEARCH ARTICLE

Transformer-Based Named Entity Recognition in Construction Supply Chain Risk Management in Australia

MILAD BAGHALZADEH SHISHEHGARKHANEH[1], ROBERT C. MOEHLER[1,2], YIHAI FANG[1], AMER A. HIJAZI[3], AND HAMED ABOUTORAB[4]

[1]Department of Civil Engineering, Faculty of Engineering, Monash University, Clayton, VIC 3800, Australia
[2]Department of Infrastructure Engineering, The University of Melbourne, Melbourne, VIC 3010, Australia
[3]Department of Civil Engineering, Al-Ahliyya Amman University, Amman 19328, Jordan
[4]School of Computing, Mathematics and Engineering, Charles Sturt University, Bathurst, NSW, Australia

Corresponding author: Robert C. Moehler (Robert.moehler@unimelb.edu.au)

This work was supported by the Monash Graduate Scholarship and Monash International Tuition Scholarship provided by Monash University, Clayton, Australia.

ABSTRACT In the Australian construction industry, effective supply chain risk management (SCRM) is critical due to its complex networks and susceptibility to various risks. This study explores the application of transformer models like BERT, RoBERTa, DistilBERT, ALBERT, and ELECTRA for Named Entity Recognition (NER) in this context. Utilizing these models, we analyzed news articles to identify and classify entities related to supply chain risks, providing insights into the vulnerabilities within this sector. Among the evaluated models, RoBERTa achieved the highest average F1 score of 0.8580, demonstrating its superior balance in precision and recall for NER in the Australian construction supply chain context. Our findings highlight the potential of NLP-driven solutions to revolutionize SCRM, particularly in geo-specific settings.

INDEX TERMS Construction supply chain risk management, named entity recognition, transformers, natural language processing, BERT.

Abbreviation

NLP	Natural Language Processing
POS	Part of Speech
NER	Named Entity Recognition
LSTM	Long Short-Term Memory
RNN	Recurrent Neural Network
BERT	Bidirectional Encoder Representations from Transformers
CSCRM	Construction Supply Chain Risk Management
GNEER	Geological News Named Entity Recognition
ELECTRA	Efficiently Learning an Encoder that Classifies Token Replacements Accurately
GPT	Generative Pre-trained Transformer
Lr	Learning Rate
W2V	Word2Vec
CBOW	Continuous Bag of Words
GS	Grid Search
NSP	Next Sentence Prediction
SG	Skip-Gram
MFT	Multi-feature Fusion Transformer

The associate editor coordinating the review of this manuscript and approving it for publication was Jolanta Mizera-Pietraszko.

I. INTRODUCTION

Natural Language Processing (NLP) is a field of computational techniques that aims to automate the analysis and representation of human language. By leveraging both theoretical principles and practical applications, NLP enables us to work with natural language data in various ways. From parsing and part-of-speech (POS) tagging to machine translation, conversation systems, and named entity recognition (NER), NLP encompasses a wide range of components and levels. It has proven itself useful in fields such as natural language understanding [1], generation [2], voice/speech recognition [3], spell correction and grammar check [4],

among others. The versatility of NLP allows it to address diverse linguistic tasks effectively [5].

The evolution of NLP can be divided into different phases that represent the progress in language generation and other language processing aspects. These phases illustrate the current state of the field, as well as ongoing trends and challenges. NLP encompasses a wide range of applications and continues to advance with computational modelling and technological innovations [6]. Furthermore, NLP involves studying mathematical and computational models related to various language aspects. It includes developing systems like spoken language interfaces that combine speech and natural language, as well as interactive interfaces for databases and knowledge bases. This enables modelling of human-human and human-machine interactions. Overall, NLP is a multidisciplinary field that intertwines computational, linguistic, and cognitive dimensions [7], [8].

A dedicated series focused on the ''Theory and Applications of Natural Language Processing'' explores the latest advancements in computational modelling and processing of speech and text in different languages and domains [9]. This highlights the rapid progress in NLP and Language Technology, driven by the increasing volume of natural language data and the evolving capabilities of machine learning and deep learning technologies. These references illustrate that NLP is a dynamic field with a solid theoretical foundation, powering numerous practical applications across various domains. NER is a method for identifying, classifying, and separating named entities into groups according to predetermined categories. NER is a crucial component of NLP technology and forms the foundation for many studies in this field. The recent advancements in deep learning have significantly improved the performance of NER applications, especially in real-world situations where high-quality annotated training data is often limited [10].

However, there has been a shift in the paradigm with the emergence of deep learning techniques driven by neural networks. These methods have shown great success in tasks such as NER, as confirmed by several studies [11]. One particularly acclaimed approach combines Long Short-Term Memory (LSTM) with CRF. In this combination, LSTM carefully captures vector representations of each word or token in a sentence, which are then fed into the CRF model for accurate sequence tagging [12]. In a groundbreaking approach, [13] combined character-level and word-level features in a hybrid network architecture. Their model utilized a BiLSTM layer followed by a log-SoftMax layer to independently decode each tag, resulting in improved accuracy. Similarly, [14] merged CRFs with information entropy, effectively identifying abbreviations of financial named entity candidates. This demonstrates the versatility of such models in specialized domains. Figure 1 shows the evolution of probabilistic models in machine learning.

To contribute to the ongoing discussion, [15] introduced a novel approach that combined a BiLSTM encoder with an incrementally decoded neural network structure. This innovative method allowed for simultaneous decoding of tags, promoting a more nuanced comprehension of textual data. Although there were various encoding strategies based on recurrent neural network (RNN) architectures, the differences in methodology became evident during the decoding phase. Recently, advanced language models like ELMo [16], GPT-4 (Generative Pre-trained Transformer) [17], and BERT (Bidirectional Encoder Representations from Transformers) [18], have emerged in the field of technology. These models have become extremely effective across various NLP tasks and have revolutionized the way we approach natural language processing. Unlike traditional methodologies that heavily relied on feature engineering, these deep neural networks possess the remarkable ability to automatically extract features from data. This characteristic has propelled them to achieve superior performance without the need for manual feature crafting or extensive domain expertise. The adoption of these sophisticated models marks a significant milestone in addressing NER tasks, facilitating more efficient and automated approaches in identifying and categorizing named entities across diverse domains and languages.

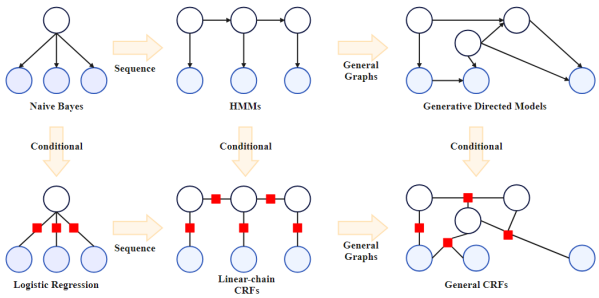


FIGURE 1. Evolution of probabilistic models in machine learning, as illustrated by the authors and inspired by [19].

The integration of NER technologies, such as advanced models like BERT, into CSCRM frameworks can greatly enhance the automatic extraction of critical information entities from a vast amount of news data. This, in turn, enables the creation of comprehensive knowledge graphs that encompass various risk factors and their potential impact on construction supply chains. These knowledge graphs hold immense value for construction firms, regulators, and other stakeholders as they foster a more resilient, transparent, and responsive ecosystem within the Australian construction supply chain. Despite its wide-ranging applications, there seems to be a dearth of research or documentation on the use of NER in construction supply chain risk management, particularly with regards to geological news in Australia [20] and [21]. This presents an opportunity for further investigation and exploration into utilizing NER to address risk management challenges within the construction supply chain domain. Specifically, it can prove valuable in leveraging geological news for more informed decision-making processes.

This research study examined the effectiveness of various BERT models in performing NER within the field of

Construction Supply Chain Risk Management (CSCRM). The primary source of information utilized is news data. This investigation breaks new ground by exploring NER applications in CSCRM specifically through the lens of news data, an area that hasn't been previously studied. The dataset consists of information gathered from multiple news outlets, providing a fertile ground for identifying and analysing numerous risk factors and how they manifest within the construction supply chain ecosystem. Through careful examination, this study uncovers several significant contributions. Firstly, it establishes a practical framework for utilizing NER to dissect real-world news data and extract valuable risk-related entities and their relationships [22]. This contributes to a deeper understanding of risk dynamics in construction supply chains. Secondly, this research provides a comparative analysis of different BERT models in accurately discerning these entities. This serves as a solid foundation for further advancements in the field. Lastly, the insights obtained through our analysis pave the way for developing more resilient and informed risk management strategies in the construction sector. It represents a significant step towards mitigating vulnerabilities within supply chains through the effective utilization of NER technologies.

II. LITERATURE REVIEW

The use of advanced language models, like BERT and GPT-3, in NER has become increasingly prevalent across various industries. From healthcare to finance, legal, and construction, businesses are leveraging these sophisticated models to accurately identify and categorize named entities within large volumes of text. These models have the remarkable ability to autonomously detect complex patterns and relationships between words without the need for labour-intensive feature engineering. This capability allows for a nuanced understanding of data, enabling critical insights extraction, better decision-making, regulatory compliance, and improved customer experiences. Additionally, advancements in transfer learning and the development of domain-specific pre-trained models have further accelerated the effectiveness and adoption of NER across diverse industries. In today's data-driven ecosystem, NER has become an indispensable tool [23], [24].

Word2Vec (W2V) has revolutionized semantic vector spaces in NLP, building on earlier foundations [25]. It introduces word embeddings through two methods: Continuous Bag-of-Words (CBOW) and Skip-Gram (SG), both sharing a neural network structure but differing in input-output management [26], [27]. Evolving beyond basic word embeddings, multi-sense and contextualized embeddings like Elmo, Bert, and Xlnet have emerged, focusing on enriched semantic understanding [28]. W2V bridges the gap between count-based models and neural networks, enhancing semantic exploration and text analytics in deep learning, thus playing a pivotal role in the evolution of pre-trained language models [29].

Reference [30] created a NER methodology to identify Chinese medicine and disease names in conversations between humans and machines. They evaluated various models, and the combination of RoBERTa with biLSTM and CRF performed the best. Using a corpus obtained through web crawling, this model achieved an impressive Precision, Recall, and F1-score of 0.96. These findings highlight its potential for enhancing medication reminders in dialogue systems. Reference [31] developed a Chinese NER model called BBIEC specifically for analysing COVID-19 epidemiological data. This model effectively processes unlabelled data at the character level, extracting global and local features using pre-trained BERT, BiLSTM, and IDCNN techniques. The BBIEC model outperforms traditional models when it comes to recognizing entities that are crucial for analysing the transmission routes and sources of the epidemic. Reference [32] proposed a BERT-Transformer-CRF based service recommendation method (BTC-SR) for enhanced chronic disease management, which initially employs a BERT-Transformer-CRF model to identify named entities in disease text data, extracts entity relationships, and integrates user implicit representation to deliver personalized service recommendations, demonstrating improved entity recognition with an F1 score of 60.15 on the CMeEE dataset and paving the way for more precise service recommendations for chronic disease patients.

Reference [33] introduced a deep learning-based Mineral Named Entity Recognition (MNER) model, utilizing BERT for mineral text word embeddings and enhancing sequence labelling accuracy by integrating the CRF algorithm's transfer matrix. Furthermore, [34] introduced a multi-task model called BERT-BiLSTM-AM-CRF. The model utilizes BERT for dynamic word vector extraction and then refines it through a BiLSTM module. After incorporating an attention mechanism network, the output is passed into a CRF layer for decoding. The authors tested the model on two Chinese datasets and observed significant improvements in F1 score compared to previous single-task models, with increases of 0.55% in MASR dataset and 3.41% in People's Daily dataset respectively. Reference [35] explored the NER task in Telugu language using various embeddings such as Word2Vec, Glove, FastText, Contextual String embedding, and BERT. Remarkably, when combining BERT embeddings with handcrafted features, the results outperformed other models significantly. The achieved F1-Score was an impressive 96.32%. Reference [36] introduced Wojood, a unique corpus specifically designed for Arabic nested NER. This corpus comprises approximately 550K tokens of Modern Standard Arabic and dialect, each manually annotated with 21 different entity types. Unlike traditional flat annotations, Wojood includes around 75K nested entities, accounting for about 22.5% of the total annotations. The accuracy and reliability of this corpus are evident in its substantial interannotator agreement, with a Cohen's Kappa score of 0.979 and an F1 score of 0.976. Furthermore, to address the limitations

of traditional methods for named entity recognition in the context of agricultural pest information extraction, [37] proposed a PBERT-BiLSTM-CRF model. This model leverages pre-trained BERT to resolve ambiguity, BiLSTM to capture long-distance dependencies, and CRF for optimal sequence annotation. The results demonstrate significant improvements in precision, recall, and an impressive F1 score of 90.24% compared to other models.

A. NAMED ENTITY RECOGNITION IN CONSTRUCTION INDUSTRY

Named entity recognition in construction has received some attention in academic literature, although the available published research in this field is relatively limited. While several studies have been conducted on this topic, the quantity of publications compared to other areas of natural language processing and construction is modest. In the realm of CSCRM in Australia, the significance of local and international news cannot be overstated.

The constantly changing geopolitical, environmental, and economic scenarios greatly impact construction supply chains. For example, the recent disruptions caused by the COVID-19 pandemic had a profound effect on the China-Australia construction supply chain. This highlighted the urgent need for timely and accurate information to effectively manage and mitigate risks [38]. The construction sector in Australia is currently facing increased supply chain risks. These risks have been amplified by the growing number of suppliers, complex work streams, stringent compliance requirements, and difficulties in finding eligible parties. It is important to note that disruptions in global supply chains, particularly those originating from regions like China, have resulted in project delays. This emphasizes the significance of international news for predicting and managing such disruptions. The lack of transparency in supply chain risk among Australian construction firms emphasizes the need for a well-informed and data-driven approach to risk management. By utilizing NER technologies, particularly in the context of geological news texts, automation can play a vital role in extracting relevant information from local and international news sources. This enhancement significantly improves the accuracy and timeliness of risk assessments and mitigating actions within the Australian construction supply chain domain.

However, the field of geological news texts is rapidly expanding, offering a wealth of valuable information. Accurately extracting this information can greatly enhance geological survey efforts. However, traditional manual extraction methods are inefficient and time-consuming, leading to lower accuracy. As the volume of geological news text data increases, these challenges become even more pronounced. It is crucial to transition towards automated extraction paradigms to address this complexity. Automating the extraction of geological news entities goes beyond just a procedural evolution; it represents a fundamental leap towards the creation of comprehensive geological knowledge graphs. These knowledge graphs can serve as structured repositories, facilitating the retrieval and analysis of geological information and propelling advancements in the field of geological surveys.

BERT, however, is a major breakthrough in the field of deep language understanding. Its architecture, which utilizes the powerful Transformer model, particularly its encoder component, has revolutionized our ability to comprehend natural language. BERT's pre-training phase involves analysing an enormous corpus of books and Wikipedia articles, allowing it to grasp the complex semantics present in textual data. The core essence of BERT lies within the encoder section of the Transformer model–an innovative design introduced by [39],–which has received widespread acclaim for its efficient parallelization of computations, greatly improving computational efficiency. The recent advancements in machine learning and NLP have significantly improved the challenges associated with manual data extraction. One notable breakthrough is the emergence of transformer-based models like BERT, which has paved the way for automating the extraction process. For example, a study introduced a method called Geological News Named Entity Recognition (GNNER) that utilizes the BERT language model to effectively extract and leverage geological data [40]. Moreover, other scholarly endeavours have demonstrated automated techniques for extracting spatiotemporal and semantic information from geological documents. These techniques are crucial for tasks such as data mining, knowledge discovery, and constructing knowledge graphs [41], [42]. The narrative above explains the importance and modern approaches used in automating the extraction of geological news information. This automation not only enhances the efficiency and accuracy of retrieving information, but it also forms a vital foundation for building comprehensive geological knowledge graphs. Table 1 compares the recent literature on NER in construction industry with current study considering their aims, models and their dataset used.

III. METHODOLOGY

A. TRANSFORMERS

As shown in Figure 2, the Transformer, distinct from traditional neural network designs, relies solely on attention mechanisms for transferring knowledge. Central to its efficiency is the attention mechanism, focusing on important data segments. Outputs are computed from weighted sums of these segments, optimizing the match between query and key. This design has significantly advanced NLP, computer vision, and spatio-temporal modeling [50], [51], enhancing model performance and training efficiency.

1) TRANSFORMER'S INPUT AND SELF-ATTENTION

The Transformer processes tokens simultaneously, adding positional encodings to each word vector. Self-attention

TABLE 1. Comparison of current study with literature.

Ref	Year	Aim	Model(s)	Dataset
[43]	2019	Extraction of geological hazard	Multi-branch BiGRU-CRF	Body of geological hazard literature and build a geological hazard knowledge graph
[44]	2019	Improving Chinese construction document	BERT	A document of Construction Management Planning and daily reports of a hospital construction project
[45]	2021	Compliance checking of the building design	GBERT	Public construction law texts
[46]	2022	Extracting defect information from noisy text in construction projects	BERT, KoBERT, KoELECTRA	Defect complaints
[47]	2022	Sort out the basic information from Chinese text about Chinese railway construction	BiLSTM	Chinese placenames and numbers
[48]	2023	Identification of building parts from Chinese construction documents	CRF-based NER model	Corpus of building's parts
[49]	2023	Compliance checking and resolve referential ambiguities	BiLSTM-CNN	Construction safety regulations
Current Study	2024	Construction Supply Chain Risk Management in Australia	Different transformer models	Unique dataset extracted from News websites

captures relationships within sequences, involving Query (Q), Key (K), and Value (V) matrices. It computes attention scores, determining focus levels on different tokens, a key factor in NLP applications [52], [53].

2) MULTI-HEAD MECHANISM AND NORMALIZATION

The multi-head mechanism assigns multiple feature expressions to each input item, enhancing the model's representational capabilities. Residual connections and layer normalization improve feature extraction and model convergence [54], [55].

$$X_{\text{att}} = X + X_{\text{att}} \tag{1}$$

$$X_{\text{att}} = \text{LayerNorm}(X_{\text{att}}) \tag{2}$$

B. BERT (BIDIRECTIONAL ENCODER REPRESENTATIONS FROM TRANSFORMERS)

BERT, based on the Transformer's encoder, represents a significant advancement in deep language comprehension. It pre-trains on vast textual data, gaining intricate semantic understanding. BERT's core relies on Transformer's encoder, praised for efficient parallel computations [39]. It processes input as token sequences, incorporating token, segment, and position embeddings for nuanced language understanding [56], [57]. Figure 3 illustrates the Transformer model's architecture with a focus on the encoder, vital for BERT.

C. KEY COMPONENTS OF BERT

1) TOKEN EMBEDDING

Assigns embeddings to tokens, representing them in high-dimensional space. Each token in BERT's vocabulary has a unique embedding learned during training [56].

2) SEGMENT EMBEDDING

Differentiates between sentences in tasks like question answering, learning separate embeddings for each sentence [56].

3) POSITION EMBEDDING

Provides positional information in the absence of a recurrent structure, crucial for understanding word order [57].

However, in addition to the BERT model, this paper also investigates other models like RoBERTa, DistilBERT, ALBERT, ELECTRA, T5, and GPT-3. Each of these models has specific adjustments and enhancements designed for different NLP tasks.

RoBERTa, also known as Robustly Optimized BERT Pre-training Approach, enhances the performance of BERT by modifying the pre-training process. This includes longer training periods, the use of larger datasets, and bigger mini-batches compared to BERT [58]. Furthermore, DistilBERT is a more compact and efficient version of BERT. It was created using a process called knowledge distillation, where the DistilBERT model learns from a pre-trained BERT model. This allows DistilBERT to maintain similar performance capabilities while being faster and more economical in terms of computational resources [59]. To make BERT more efficient, ALBERT (A Lite BERT) utilizes techniques like factorised embedding parameterisation and cross-layer parameter sharing. These methods reduce the size and increase the speed of BERT [60]]. Instead of using the masked language modelling objective like BERT, ELECTRA (Efficiently Learning an Encoder that Classifies Token Replacements Accurately) takes a different approach. It utilizes a pre-training task called replaced token detection, which aims to achieve more efficient pre-training [61]. T5 (Text-To-Text Transfer Transformer) takes a distinctive approach by transforming every NLP problem into a text-to-text format. This simplifies the application of the model to various NLP tasks [62]. Finally, the GPT model is a ground-breaking technology that has revolutionised NLP. Through unsupervised learning on vast amounts of text data, it has successfully generated text that closely resembles human writing [63]. GPT-3 is the successor to GPT-2 and boasts a significant raise in both parameter count (from 1.5 billion to 175 billion) and data processed (from 40 GB to 45 TB), making it the largest language model ever created [64].

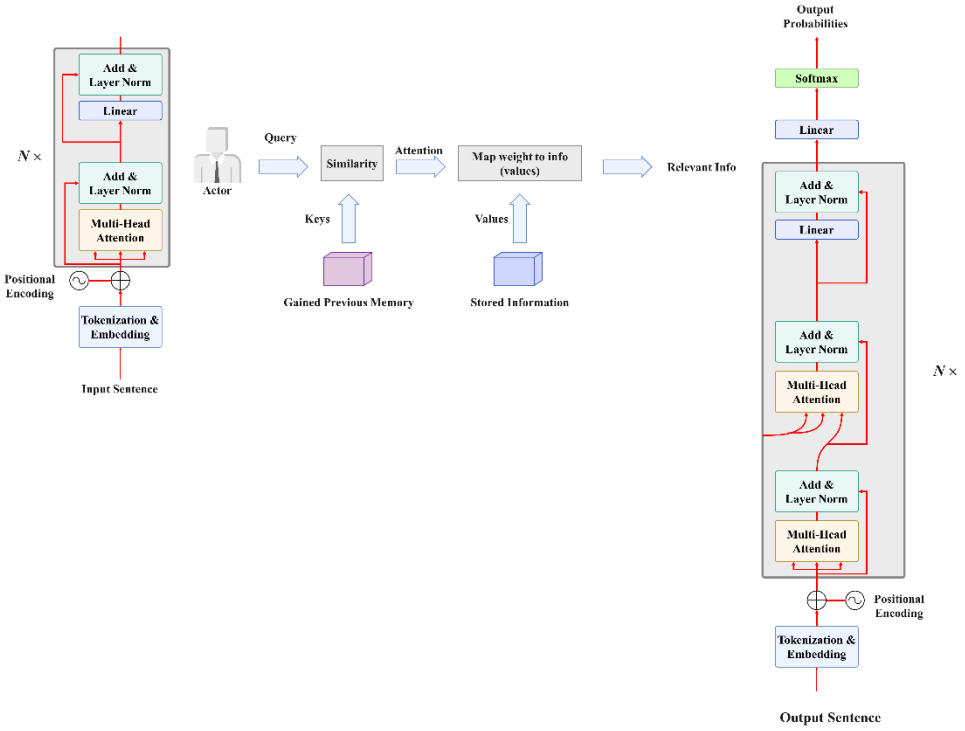


FIGURE 2. Transformer's architecture, as illustrated by the authors and inspired by [39].

All models were obtained from the Hugging Face model repository. The selection of these models was predicated on their proven efficacy in the field of NLP and their suitability for the task at hand, as evidenced by their widespread use and robust performance across numerous benchmarks. Each model was meticulously evaluated to ascertain its compatibility with the specific requirements of recognizing entities within the context of Australian construction supply chain risk management. This involved an extensive comparative analysis to identify the model that demonstrated the highest accuracy and efficiency when applied to our annotated dataset. However, we continuously monitored the Hugging Face model hub for the latest developments in GPT models. However, as of the completion of our research, GPT-4 had not been released or made publicly accessible in this repository or any other known open-source platform.

D. DATA GATHERING

This study conducted a thorough investigation to create a detailed risk categorisation specifically for managing risks in the construction supply chain in Australia. This involved carefully reviewing existing literature and incorporating insights from the Cambridge Taxonomy of Business Risks [65]. As shown in Figure 3, the resulting risk categorisation covers a wide range of risks commonly found in the Australian construction supply chain, providing a strong basis for the following stages of this study.

After establishing the risk taxonomy, the attention turned to collecting a comprehensive dataset for thorough analysis of the identified risks. A specialised News API was utilised to search through approximately 2000 articles from renowned news sources like *The Australian, Sky News Australia, Bloomberg, CNN, Reuters,* and *Google News*. The careful selection of news sources is paramount to ensure the comprehensiveness, diversity, and reliability of the dataset. Guided by specific criteria aimed at capturing a wide range of perspectives and factual reporting, the chosen sources each boast a longstanding reputation for journalistic integrity and reliability. This approach mitigates biases and inaccuracies in data-driven research, as emphasized by [66]. To ensure a balanced representation of global news, sources from various geographical locations and political orientations were selected, aligning with the recommendations for achieving diversity in news recommendation systems highlighted by [67]. These sources are known for their comprehensive coverage across a range of topics, including politics, economics, and international affairs, critical for a well-rounded dataset. Furthermore, the importance of digital accessibility and archival features was a key factor, ensuring ease of data collection and analysis. Each selected source, including The Australian and Sky News Australia, offers unique insights into the Asia-Pacific region; Bloomberg is revered for its financial news; CNN and Reuters provide extensive global coverage; and Google News aggregates a diverse range of sources, enabling a comprehensive topical analysis. This data

FIGURE 3. Construction supply chain risk taxonomy used in this research.

collection approach was carefully designed to adhere to web scraping guidelines and ensure ethical acquisition of data. The result was a diverse and extensive dataset that provided ample material for empirical investigation of the specified risks within the CSCRM domain using NER.

1) ANNOTATION OF TEXT CORPUS

For dataset annotation, sequence labelling is a critical step that helps organise data for further analysis. Among various labelling methods used in scholarly research, this study utilizes the ''BIO'' labelling scheme due to its effectiveness and widespread acceptance. This labelling convention, commonly employed in NER, offers a systematic approach to annotate text sequences, allowing for a detailed understanding of the text structure. The ''BIO'' labelling scheme consists of three annotations: ''B-X,'' ''I-X,'' and ''O.'' In this scheme, the letter ''B'' indicates the beginning of a named entity in the text. The letter ''I'' represents the middle and concluding segments of the named entity. Lastly, the letter ''O'' denotes text segments that do not contain a named entity [68].

After carefully applying the ''BIO'' labelling scheme to the news texts, we were able to obtain a substantial data-set with labelled information. Our statistical analysis after labelling revealed an impressive count of 39,500 entities across six different categories, as shown in Table 2. Figure 1 shows the methodology of the current research work. Initially, news articles are gathered through a web crawler and News API, systematically aggregating them into a news database. This corpus of text is then subject to data cleaning and annotation processes, during which irrelevant information is purged and relevant entities are marked according to predefined categories such as ''Recession'' (RRE) and ''China'' (GPU). Following this preparatory phase, the text is tokenized and masked in preparation for Named Entity Recognition (NER), a crucial step for understanding the context and extracting meaningful information. Various Transformer-based NER models, including BERT, RoBERTa, DistilBERT, ALBERT, ELECTRA, T5, and GPT-3, are evaluated to determine the most effective framework for the task. The chosen model is then trained, tested, and validated using the annotated datasets. The annotated entities in the example news article illustrate the application of the methodology, with entities such as ''Recession'' and ''Interest rate'' tagged as RRE, indicating risk events, while ''China'' and ''International Monetary Fund'' are tagged as GPU and OSC, respectively, highlighting geopolitical units and organizations.

TABLE 2. Annotated corpus's entities number.

Entity	Category	Occurrences	Examples
PER	Person's names	560	John Smith, Angela Hughes
RRE	The most relevant risk events from risk taxonomy	3674	Inflation, Labour Strike
PNR	Political, Nationalities, and Religious groups	570	Labor Party, Australian nationals
OSC	Organisations, Suppliers, and Companies	1416	BHP Billiton, Sydney Constructions Pty Ltd
GPU	Geo Political Units	3606	New South Wales, Victoria
CMS	Construction Materials	570	Steel, Concrete

2) IAA ANALYSIS FOR ANNOTATED CORPUS IN CSCRM

This section presents the methodology and findings of the Inter-Annotator Agreement (IAA) analysis conducted to ensure the accuracy and reliability of our NER dataset annotations. Given the critical role of high-quality annotations in

NER research, especially within specialized domains such as CSCRM, assessing annotation consistency is paramount [69].

Initially, the paper's authors annotated the entire dataset, ensuring a preliminary layer of domain-specific insights and linguistic accuracy. Subsequently, a specialized validation team–comprising two experts in construction management, one in linguistics, and two in computational linguistics–conducted a thorough review of these annotations. This team's primary role was to validate the accuracy, relevance, and computational suitability of the initial annotations. Following this validation phase, statistical analyses, including Inter-Annotator Agreement metrics, were calculated to quantify the consistency and reliability of the annotations, ensuring the dataset's robustness for NER tasks within CSCRM.

Cohen's Kappa (κ) measures the agreement between two annotators on a categorical scale. It is calculated as:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{3}$$

where p_o is the observed agreement proportion, and p_e is the proportion of agreement expected by chance. Observed agreement (p_o) is calculated by dividing the number of instances both annotators agree on by the total number of instances. Expected agreement (p_e) is based on the probability that annotators randomly agree, considering the distribution of each category.

Fleiss' Kappa (κ) assesses agreement among three or more annotators and is calculated with:

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} \tag{4}$$

where $\bar{P}$ is the average proportion of agreement observed between all pairs of annotators, and $\bar{P}_e$ is the expected agreement by chance. Calculating $\bar{P}$ involves averaging the proportions of agreement across all annotators for each item, while $\bar{P}_e$ considers the chance agreement across all categories. Table 3 shows the results of IAA in this study.

TABLE 3. Inter-annotator agreement results for entity categories.

Entity	Cohen's Kappa (κ)	Fleiss' Kappa (κ)
PER	0.90	-
RRE	0.85	-
PNR	0.80	-
OSC	0.88	-
GPU	0.92	-
CMS	0.87	-
Overall Fleiss' Kappa		0.86

The IAA results, featuring both Cohen's and Fleiss' Kappa scores, indicate a high degree of consistency among annotators across various entity categories relevant to construction supply chain risk management. Specifically, Cohen's Kappa scores range from 0.80 to 0.92, reflecting substantial to almost perfect agreement on entities from personal names to geopolitical units and construction materials. The overall Fleiss' Kappa score of 0.86 further underscores a strong consensus among all annotators, affirming the dataset's reliability and the effectiveness of the multidisciplinary annotation approach.

3) DATA PREPARATION

In the division of our dataset into training, validation, and test sets, we implemented a stratified shuffle split approach to maintain the distribution of labels consistent across all subsets. The stratification process ensured that each subset was representative of the full dataset, with all categories of labels proportionally reflected in the training, validation, and test sets. This method is particularly important in our context to avoid skewed or biased model training and evaluation, given the uneven distribution of risk-related entities in construction supply chain management articles. Before the split, the dataset was shuffled to guarantee the randomness of data distribution, thereby preventing any potential order effects that could influence the model's learning pattern. The shuffling and stratification were performed using robust functionalities provided by data processing libraries in Python, ensuring reproducibility and adherence to standard data preparation practices in machine learning.

In order to train the transformers models, it requires powerful computational resources. Table 4 provides detailed information about the hardware and software used in this experiment, giving a comprehensive understanding of the infrastructure that supported the training of the transformer models in this study.

TABLE 4. Experimental setup of this study.

Type	Configuration	Features
Software	CUDA	11.5
	Python	3.9
	Numpy	1.21.2
	Scikit-learn	0.24.2
	Pandas	1.3.3
	TensorFlow	2.6
	PyTorch	1.9.0
Hardware	Operating System	CentOS Version 8.4
	Video RAM	11 Gigabytes GDDR6
	RAM	32.0 Gigabytes
	Processor	Intel Core i7-12700, 12th Generation

TABLE 5. Models' hyperparameters.

Hyperparameters	Values
Drop Rate	0.50
lr	3e-5
Batch Size	32
Epochs	10
Max len	128

During the model training phase, the choice of hyper-parameters greatly affects the outcomes. To ensure consistency and reduce variability between experiments, this study used a fixed set of hyper-parameters for training different models. The important parameters involved in the training process are listed in Table 3. In this table, an epoch refers

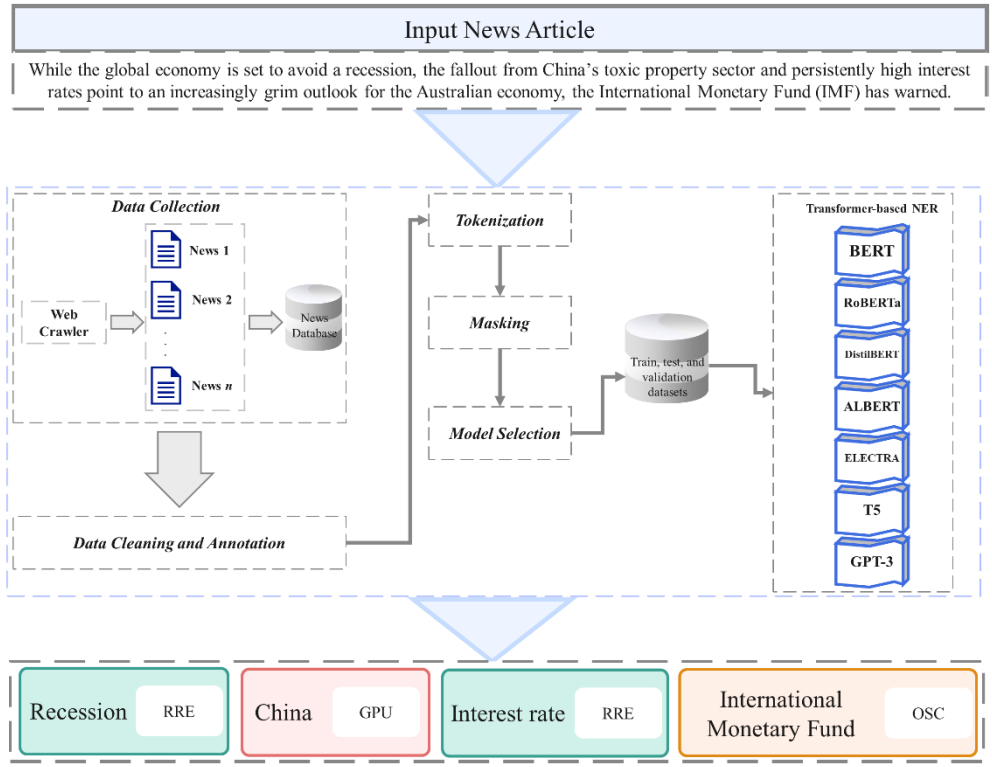


FIGURE 4. Methodology of this research.

to one complete iteration over the entire training dataset, max len indicates the maximum sequence length, batch size determines how much data is processed in each training iteration, lr controls the rate of learning, and drop rate helps prevent over-fitting in the neural network.

The training phase commenced with the initialization of model weights, often starting from pre-trained checkpoints available through the Hugging Face repository, which provided a foundation of linguistic knowledge beneficial for downstream tasks. The data was then pre-processed to align with the input requirements of the transformer models, including tokenization, encoding of the BIO tags, and application of the 'max len' parameter to standardize sequence lengths. Each word in the corpus was converted into tokens, and special tokens were added where necessary to signify the beginning, separation, and end of sentences, as required by the model architectures.

For each epoch, the model processed the training data in batches, with the 'batch size' calibrated to ensure efficient memory utilization and gradient updates. The training loop involved forward propagation of the batch through the model, calculation of the loss function comparing the predicted entity tags with the correct tags, and back-propagation to adjust the weights of the model using the learning rate ('*lr*') as a factor in the weight update rule. The 'drop rate' was strategically applied within the model's layers to randomly deactivate certain neurons, thereby encouraging the model to learn more robust features that do not rely on any small set of neurons.

Model validation occurred at the end of each epoch, where a subset of the data, held out from the training set, was used to evaluate the model's performance. This step was critical to monitor for over-fitting and to tune the hyper-parameters if necessary. The model with the best validation performance was then selected for final evaluation. In the testing phase, the model's generalization capabilities were rigorously assessed using a distinct set of annotated news articles. This phase was designed to simulate the model's deployment in real-world scenarios, where it would encounter data variations and complexities. The performance metrics–precision, recall, and the F1-score–were meticulously calculated, providing quantitative insights into the model's accuracy, its ability to detect relevant entities (true positives), and its precision in not misclassifying non-relevant elements as entities (true negatives).

For our study, we carefully divided the labelled data-set into three separate subsets: the training set, validation set, and test set. We followed a distribution ratio of 8:1:1 to allocate the entities. This means that we had 31,600 entities for training, 3,950 entities for validation, and another 3,950 entities for testing. This division is crucial in order to train and evaluate models effectively and ensure their robustness and ability to generalise in line with suggestions from [70]. To study NER in CSCRM, we trained seven different models using a designated training data-set. After training, we evaluated the performance and effectiveness of these models on a separate test set for NER tasks. This

approach is similar to the method used by when evaluating multiple models to determine the best one for NER tasks in a similar domain [46].

IV. RESULTS AND DISCUSSION

This section discusses the results of different models that were used for NER in news articles focusing on construction supply chain risk management.

A. RESPONSIBLE AI CONSIDERATIONS IN TRANSFORMER MODELS FOR NER

In the field of NER, it is important to explore the ethical implications and responsible AI considerations that emerge when using transformer models like BERT, RoBERTa, DistilBERT, ALBERT, ELECTRA, T5, and GPT-3. While these models have good numbers on their side, there are other factors beyond their accuracy scores that reveal some ethical problems associated with them including biases, data privacy, transparency, and interpretability among others [71]. The most important thing is to ensure that predictions by the model are faithful and correspond with what the input provided says as well as the world outside. Ensuring adherence to fidelity is one of the significant ways to determine whether a transformer model can correctly identify named entities across varying contexts and subtle language differentiations. This means that thorough validation is needed especially in multi-lingual or domain-specific NER applications to ensure dependable outputs that are contextually appropriate. Equally important for responsible AI discussions is Monotonicity another often overlooked attribute. Trustworthiness requires assurance that small changes in input will result in predictable small changes in output; otherwise, it would not be reliable. Assessing how these models maintain monotonicity, especially in scenarios where small alterations in input should not significantly impact NER predictions, serves as a benchmark for their robustness and responsible usage.

Other key factors are identification and mitigation of biases. The data they are trained on is varied and extensive, meaning that these models are likely to inherit the societal biases present in that data, which then causes distorted predictions. Biased systems may have different accuracy rates across different racial groups or reproduce stereotypes. This section looks at how each model deals with bias through adjusting the algorithm or curation of datasets, thereby showing its dedication towards ensuring equity in NER tasks. Also, transparentness and interpretability are crucial for ensuring ethical deployment of AI. Trust building comes from the capability to explain model predictions as well as comprehend why a particular decision-making process is used. It is important to establish how these transformer models enable interpretation and whether they make clear case for predictions in order foster responsible use of AI in the NER area.

B. EXPERIMENT DESIGN AND ASSESSMENT

When evaluating the performance of different models in NER, precision (P), recall (R), and F1-score (F1) are commonly used metrics. These metrics help assess how well the models perform. The specific computational formulas for these metrics are as follows:

$$P = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{5}$$

$$R = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{6}$$

$$F1 = 2 \times \frac{P \cdot R}{P + R} \tag{7}$$

where *TP* represents the number of correctly identified entities or true positives. *FP* denotes the number of incorrectly identified entities or false positives. Lastly, *FN* signifies the number of missed entities or false negatives. Precision, Recall, and F1 scores of the mentioned models are shown in Figure 5.

C. MODEL EVALUATION AND COMPARISON

In the provided evaluation metrics (Precision, Recall, and F1 Score), Table 4 presents a comparative analysis of the models. RoBERTa stands out with an impressive average F1 score of 0.8580, which indicates a well-rounded performance in both precision (0.9341) and recall (0.8023). On the other hand, T5 exhibits the highest average precision value of 0.9924 but suffers from a low recall of 0.3645, resulting in a modest F1 score of 0.5115. These differences highlight varying capabilities among the models in accurately identifying entities and retrieving relevant instances from the news dataset.

When evaluating the performance of models in various categories such as PER, RRE, PNR, OSC, GPU, and CMS, it becomes evident that each model has its strengths and weaknesses. In terms of precision, almost all models demonstrate high accuracy in the PNR and CMS entities. Some even achieve a perfect score of 1.0000. However, the OSC entity poses challenges for all models. Though T5 exhibits the highest precision score of 1.0000 in this category, its recall rate is notably lower. This suggests that factors like entity characteristics or variations in training data quality and quantity significantly impact the overall performance of these models across different categories. The performance of models in NER tasks is significantly influenced by their underlying architectures and training data. Transformer-based models like BERT, RoBERTa, and DistilBERT excel in capturing contextual relationships, which are crucial for NER tasks. On the other hand, models like T5 and GPT-3 approach NER differently as text-to-text and generative models, respectively.

GPT-3's performance in NER tasks is generally lower than supervised baselines due to the inherent gap between NER (a sequence labelling task) and GPT-3's nature as a text generation model. However, adaptations such as GPT-NER have been proposed to bridge this gap by transforming the

sequence labelling task into a generation task that can be easily tailored for large language models like GPT-3 [60]. Moreover, ELECTRA uses a unique pre-training task where token detection is replaced with distinguishing ''real'' from ''fake'' input data. This can potentially improve its NER performance by reducing false positives and negatives in entity recognition [72]. When evaluating and selecting models for implementation within the construction supply chain domain, it is crucial to consider both the architectural differences of the models and the nature of the NER tasks. This analysis highlights the significance of this dual consideration.

D. COMPARISON WITH LITERATURE

This section compares the results of this research with recent research works from literature, considering their performances and F1 scores. The transformative impact of transformer models on NER is evident across a range of linguistic tasks, as seen in recent studies. [73] compared BERT, RoBERTa, and XLNet to non-transformer models, finding that the transformer models consistently outperformed others in terms of the F1 score, regardless of domain. This finding is reflected in the current data, which showcases RoBERTa's leading performance with an average F1 score of 0.8580.

The work of [74] provides an interesting perspective by evaluating DistilBERT's performance on medical texts. The findings revealed that while DistilBERT achieves F1 scores comparable to those of BERT models on medical texts, its efficiency in runtime and resource usage stands out. It is suggested that DistilBERT achieves a balance between performance and operational efficiency, with an average F1 score of 0.8240, which may suggest some trade-offs in generalization capacity. Reference [75] further underscore the capability of BERT Multilingual Cased and Uncased models, achieving average F1 scores of 85.41 and 83.52, which suggests that multilingual models retain high performance even in the context of Indonesian health insurance data. An additional layer of insight is provided by [76], who explored the use of machine translation to generate Persian named entity datasets, achieving an impressive F1 score of 85.11, highlighting the potential of transformer models in low-resource languages. Reference [77] reported that the MuRIL (Large) model performs well in multilingual NER tasks for Hindi and Bangla, with F1 scores of 0.69 and 0.59, respectively, indicating the adaptability of transformer models across different languages.

Moreover, the research by [78] on a transformer-based system for English NER achieved a macro F1 score of 72.50 on a test set, reinforcing the effectiveness of these models in complex NER tasks. Reference [79] demonstrated the utility of the T5 and transformer encoder for multilingual complex NER, showing an increased model F1-score by 4% in English when subtoken checks were introduced. Reference [80] proposed a cost-sensitive contextualized model for Bangla NER, indicating the need for specialized approaches for low-resource languages, achieving an F1 Macro score of 65.96. Reference [81] introduced BioNerFlair, outperforming state-of-the-art models with an F1-score of 90.17, suggesting that alternative architectures could offer competitive results in domain-specific tasks. Reference [82] demonstrated that ensembles of transformer-based models could further improve NER results, achieving a 76.36 F1-score.

These studies collectively illustrate the broad applicability and high performance of transformer models in NER tasks across languages and domains. They provide a strong foundation for further refinement and application of these models in increasingly diverse linguistic environments. The comprehensive analysis across these works underscores the robustness of transformer architectures, particularly RoBERTa, in handling the complexities inherent in NER tasks, as well as the nuanced capabilities of models like DistilBERT and BERT Multilingual in specific contexts. Continuing the exploration of transformer model capabilities, [83] introduced the Multi-feature Fusion Transformer (MFT) for Chinese NER, significantly outperforming mainstream models with an F1 score of 95.77 on the Resume dataset, demonstrating the potential of incorporating character radical information in enhancing semantic understanding. This underscores the adaptability of transformer models to the intricacies of logographic languages.

Reference [84] conducted a comparative study of pre-trained language models for NER in clinical trial eligibility criteria, finding that domain-specific transformer models like PubMedBERT outperformed general transformer models, indicating the importance of domain-specific pre-training for achieving high F1 scores in specialized tasks. Reference [85] demonstrated that multi-task learning using transformer models could improve the F1 score by 2% in biological named entity recognition tasks, highlighting the benefits of joint training on related NLP tasks. In the medical domain, [86] employed transformer encoder and pre-trained language models for Chinese medical NER, achieving an F1 score of 82.72, showcasing the effectiveness of model ensembles in technical fields. Reference [87] used a multi-embeddings approach coupled with deep learning for Arabic NER, achieving an F1 score of 77.62 on the AQMAR dataset, thus setting a new performance standard for NER in Arabic. Reference [88] utilized the transformer model XLM-Roberta for Hindi NER, achieving F1-scores of 0.96 (micro) and 0.80 (macro), further validating the effectiveness of transformers in language-specific NER tasks.

These studies not only reinforce the state-of-the-art status of transformer models in NER but also illustrate the ongoing innovations aimed at optimizing their performance and expanding their applicability across languages and specialized domains.

E. RELATIONSHIP BETWEEN ENTITY CATEGORIES AND THEIR AMOUNTS

As shown in Figure 4, the performance of different models in identifying entities varies significantly, especially when

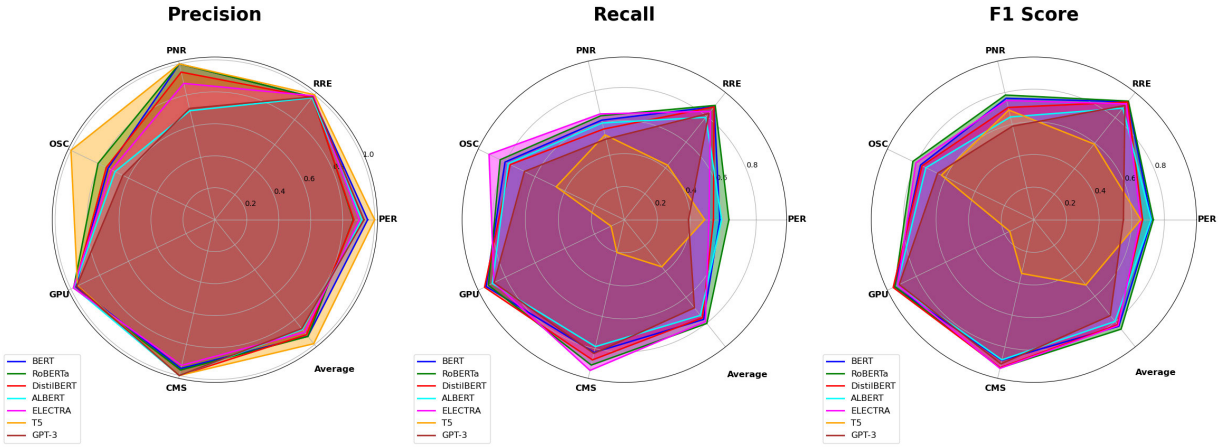


FIGURE 5. Statistical results of different models for NER in CSCRM.

compared to the frequency of occurrence for each entity type. Entities that occur more frequently, such as RRE (3674) and GPU (3606), generally have higher precision and recall scores across most models. This indicates that having a larger dataset contributes to better model performance. For example, models like BERT, RoBERTa, and ELECTRA show notably high F1 scores for entities like RRE and GPU. In a research paper, when examining various transformer models, including BERT, RoBERTa, and ELECTRA using a detailed emotion, it was found that the size of the model did not have a significant impact on the task of emotion recognition. This suggests that while data size may affect performance, the size and architecture of the models also play important roles [89].

However, there are some exceptions to this pattern. Despite having an equal number of occurrences in the data-set (570), both PNR and CMS exhibit fluctuating performance between categories. This variability suggests that the quantity alone is not the sole factor determining model performance; the complexity or uniqueness of the entity type may also play a role. For example, a comparative analysis also examined how well these models recognise emotions from texts, providing further insight into their performance on entity recognition tasks across different categories and data-sets. Additionally, a separate study focused on domain specific applications explored these models' ability to extract various clinical concepts, offering insights into their capacity to handle different types of entities and understanding how domain and data-set size can impact model performance [90].

RoBERTa consistently achieves the highest F1 score among the models, closely followed by BERT. This indicates that having a large amount of data can improve model performance, but the architecture and training techniques are still crucial factors. When it comes to tasks requiring a balanced precision and recall, RoBERTa or BERT are considered the most suitable options. Furthermore, a study on recognizing Protected Health Information (PHI) entities revealed differences in training times between these models, which could indirectly impact their performance on entity recognition tasks. These findings suggest that training time and computational resources may also influence how well different models perform in entity recognition tasks [91]. Furthermore, RoBERTa differs from BERT in several key aspects, including removing the Next Sentence Prediction (NSP) task, training with larger mini-batches and learning rates, and using longer sequences. These changes enable RoBERTa to better capture context and semantics, crucial for tasks like NER that require understanding of complex sentence structures and subtle language cues. Furthermore, RoBERTa is trained on a much larger corpus than BERT, which allows it to develop a more nuanced understanding of language. These modifications contribute to RoBERTa's superior performance in scenarios demanding high precision and recall, making it a more suitable option for balanced performance in NER tasks [89].

When using these models, it's important to consider both the frequency of the entity in the data and its complexity to ensure the best results. Research on NER using RoBERTa and ELECTRA models has shown that performance varies depending on the specific model and dataset used. For instance, an ELECTRA-based model performed better than BERT-based models when working with a dataset related to drugs, as measured by its F1 score. This highlights how the choice of model and characteristics of the dataset significantly impact entity recognition performance [92]. It underscores the importance of considering factors such as data availability, model architecture, and entity complexity in order to achieve optimal results in entity recognition tasks.

F. IMPACT OF HYPER-PARAMETER FINE TUNING ON MODELS PERFORMANCE

Grid search (GS) is a traditional technique used in fine-tuning parameters in machine learning and deep learning tasks, including NLP with popular models such as BERT. It systematically explores a range of parameter options, typically in a

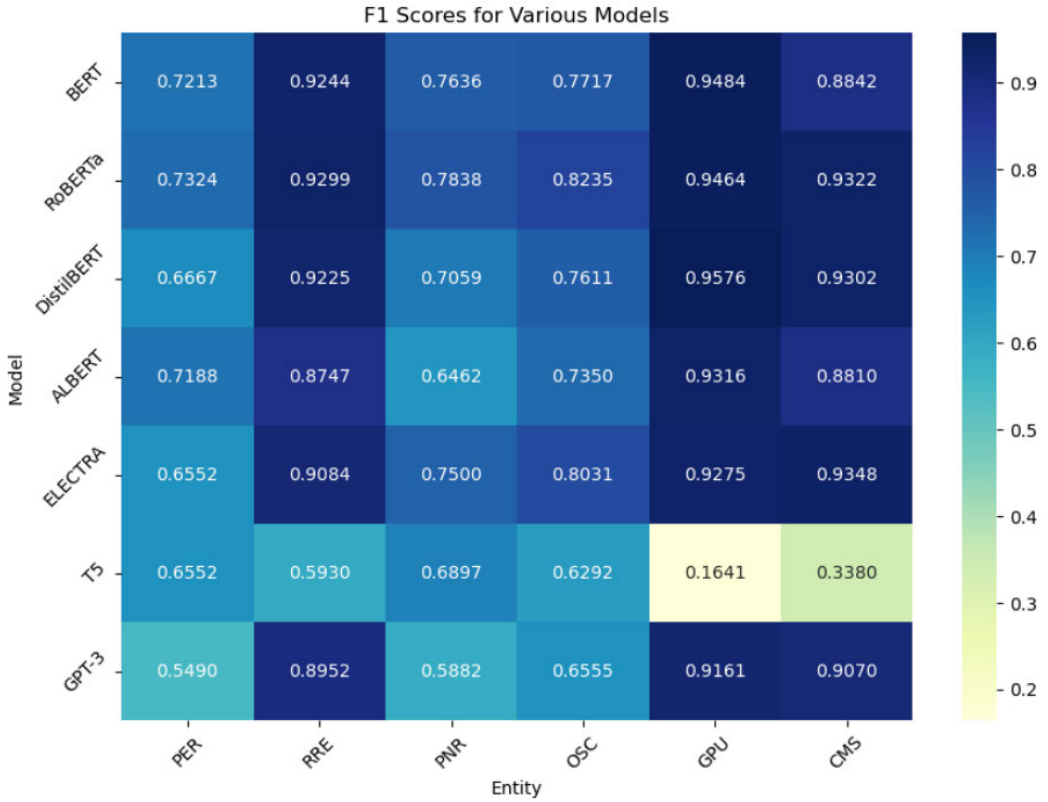


FIGURE 6. F1 scores of different models for each entity.

methodical grid-like pattern, to determine the most effective parameters for a specific model. The strength of GS lies in its simplicity and thoroughness, ensuring that each combination of parameters is evaluated to find the optimum setting. This exhaustive approach, while computationally expensive, provides a comprehensive survey of the parameter space, often leading to more reliable and robust model performance. Studies have shown that GS can be more efficient and produce better uniformity in parameter selection compared to random or heuristic methods, making it a reliable choice for achieving high accuracy and generalization in machine learning models [93]. While GS can be time-consuming and may require considerable computational resources, its effectiveness in identifying optimal parameters makes it a valuable tool in the machine learning toolbox, especially for tasks requiring high precision.

For instance, in a research, the authors used grid search to thoroughly refine BERT and other models using the DuoRC dataset. They focused on key hyper-parameters such as maximum sequence length, maximum question length, document stride, and training batch size, tweaking them prior to training to enhance model performance [94]. Also, in common practice, GS is performed across a range of parameter values to figure out the combination that yields the best results for a given task. This approach plays an especially crucial role in the NLP field, where parameters like learning rate, batch size, and optimizer type can greatly affect the performance of models like BERT. The GS method works by thoroughly assessing models across a particular parameter grid, set up as follows:

$$G = \{(p_1, p_2, \ldots, p_n) \mid p_i \in P_i\} \tag{8}$$

where P_i shows the set of possible values for hyper-parameter i.

This study involves exploring different sets of hyper-parameters, including learning rates (lr), batch size (BS), epsilon (ε), and two unique optimizers–Adam and AdamW, as shown in Table 5. The set P_i represents the set of possible values for hyper-parameter i. This section specifically looks at how these factors affect the overall performance of models in NER tasks in construction supply chain risk management. This is distinct from the previous section, which evaluated the performance of models on individual entities.

TABLE 6. Hyperparameter sets.

Hyperparameter	Values
Learning Rate	1e-6, 5e-6, 1e-5, 3e-5, 5e-5, 1e-4, 5e-4, 1e-3
Epsilon	1e-7, 1e-8, 1e-9
Batch Sizes	16, 32, 64
Optimizers	Adam, AdamW

TABLE 7. Best hyper-parameter combinations based on GS.

Model	Hyper-parameters				Precision	Recall	F1-score	Efficiency (s)
	lr	BS	ε	Optimizer				
BERT	3e-05	16	1e-9	Adam	0.7882	0.8449	0.8097	161.2068
	1e-4	32	1e-9	Adam	0.8032	0.7797	0.7824	145.9247
	3e-05	16	1e-9	Adam	0.7882	0.8449	0.8097	161.2068
	1e-3	64	1e-9	Adam	0.1372	0.1428	0.1399	134.9384
RoBERTa	1e-3	64	1e-9	Adam	0.7753	0.8350	0.7944	106.7353
	1e-3	64	1e-9	Adam	0.7753	0.8350	0.7944	106.7353
	3e-05	32	1e-8	AdamW	0.7618	0.8412	0.7903	116.7825
	1e-06	64	1e-8	Adam	0.6850	0.7692	0.7138	106.5951
DistilBERT	1e-3	16	1e-7	AdamW	0.7898	0.7963	0.7841	84.8277
	5e-05	32	1e-9	AdamW	0.7985	0.7787	0.7725	75.5385
	1e-3	64	1e-8	AdamW	0.7351	0.7998	0.7569	69.8459
	1e-3	64	1e-9	Adam	0.1378	0.1428	0.1403	68.8453
ALBERT	3e-5	32	1e-8	AdamW	0.8100	0.8029	0.8018	155.8440
	1e-3	16	1e-9	Adam	0.8235	0.7424	0.7738	163.4706
	5e-5	32	1e-7	AdamW	0.7900	0.83027	0.7994	155.8783
	1e-3	64	1e-9	Adam	0.1378	0.14285	0.14032	147.5978
ELECTRA	1e-3	32	1e-9	AdamW	0.7910	0.8054	0.7933	149.6477
	1e-3	32	1e-9	AdamW	0.7910	0.8054	0.7933	149.6477
	5e-5	16	1e-8	AdamW	0.7480	0.8201	0.7766	165.5166
	1e-3	64	1e-9	Adam	0.1378	0.1428	0.1403	137.6781

Using GS, this study found 144 unique combinations to assess how different mixes of hyper-parameters affect model performance in NER tasks. Table 6 shows the results of these assessments, pointing out the best combination for higher precision, recall, and F1 score, as well as the most efficient combination. It also considers the less successful combinations, giving a full view of performance across various hyper-parameter setups. When conducting hyper-parameter tuning, it's essential to focus on models that are most relevant and promising for the task at hand. In this context, we concentrated on transformer models like BERT, RoBERTa, DistilBERT, ALBERT, and ELECTRA, excluding T5 and GPT-3. This decision was based on several key considerations. First, transformer models have shown exceptional performance in understanding context and generating language, making them ideal for a wide range of NLP tasks. Each of these models, from BERT's pioneering architecture to ELECTRA's efficiency in understanding language, has unique strengths that make them suitable for in-depth hyper-parameter optimization. Second, due to their architecture and training methods, models like T5 and GPT-3 require significantly more computational resources for training and tuning, which may not be feasible or necessary for the specific objectives of our project [95]. Moreover, GPT-3's closed-source nature and licensing limitations also posed a constraint [96]. Therefore, our focus on the selected transformer models was driven by a balance of performance, resource availability, and specific model characteristics that align with our project goals. The best hyperparameter combinations are shown in Table 6.

In the conducted grid search for named entity recognition within the context of Australian construction supply chain risk management, distinct trends and implications have been revealed through the comparative analysis of models such as BERT, RoBERTa, DistilBERT, ALBERT, and ELECTRA, in relation to various hyper-parameters and optimizers. It has been observed that competitive performance is exhibited by both BERT and RoBERTa, with BERT slightly outperforming in terms of recall, indicative of its effectiveness in identifying relevant entities. Conversely, RoBERTa is distinguished by offering a more balanced trade-off between precision and recall, coupled with higher efficiency, positioning it as a time-efficient alternative.

DistilBERT, characterized by its lighter architecture, has been noted for its efficiency, achieving this without significant sacrifices in precision and recall, thereby emerging as a robust option under constraints of computational resources or time. In a different vein, ALBERT has been recognized for its precision, especially under specific hyper-parameter configurations, rendering it particularly suitable for tasks where precise identification is critical. ELECTRA, while not outshining in specific metrics, has been acknowledged for providing a consistent balance across various performance measures, which can be advantageous in scenarios demanding uniform performance.

Further insights have been gained into the effects of hyper-parameters and optimizers, where the learning rate has been identified as a critical factor influencing model performance. Generally, lower learning rates have been found to yield better recall and F1-scores, suggesting the benefit of a cautious approach in weight updating within this specific domain. However, it has been noted that excessively low learning rates might impair the learning capabilities of the model. Consistently, larger batch sizes have been associated with diminished performance, indicating the effectiveness of smaller batch sizes for this particular application.

Regarding the choice of optimizer, no consistent preference has been discerned between Adam and AdamW. However, it has been observed that models employing

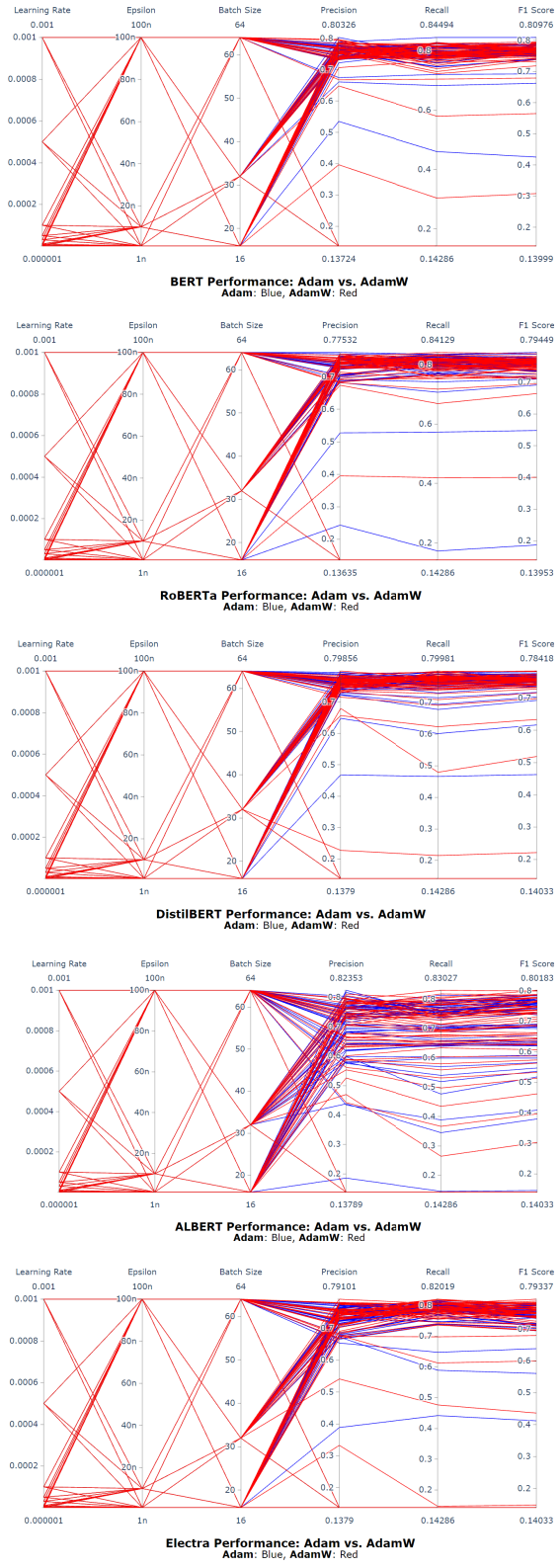


FIGURE 7. Comparative analysis of transformer models' performance using adam and adamw optimizers across various hyperparameters.

AdamW, particularly in the cases of DistilBERT and ALBERT, demonstrate enhanced efficiency. This improved efficiency might be attributable to the weight decay strategy of AdamW, which aids in expediting the convergence process. The findings underscore the necessity for a tailored selection of models and hyper-parameters in named entity recognition tasks, with the aim of aligning them with the specific requirements of the task at hand. BERT and RoBERTa have been noted for their proficiency in recall, while DistilBERT and ALBERT excel in efficiency and precision, respectively. ELECTRA, as a model, stands out for its well-rounded performance. The employment of lower learning rates and smaller batch sizes has generally been found to be more effective, while the choice between Adam and AdamW optimizers appears to be more influenced by considerations of efficiency than by factors of precision, recall, or F1-score. Figure 5 shows comparative analysis of transformer models' performance using Adam and AdamW optimizers across various hyper-parameters.

When assessing precision, recall, and F1 scores in relation to optimizer choice, it is apparent that AdamW tends to enhance model performance across most models, suggesting its superiority in handling weight decay and perhaps aiding in generalization. However, the degree of this enhancement varies, indicating differing levels of sensitivity among the models to the optimization method. Considering learning rates, a lower learning rate coupled with AdamW optimizer seems to consistently benefit models like BERT, ALBERT, and to some extent, RoBERTa, in achieving higher F1 scores. DistilBERT and Electra, on the other hand, exhibit a less pronounced preference, indicating a potential robustness to a wider range of learning rates or an architecture that is less amenable to the subtle improvements offered by AdamW's weight decay.

The impact of batch size on model performance, particularly with BERT and ALBERT models using the AdamW optimizer, reveals nuanced insights into their learning dynamics. Larger batch sizes, when combined with AdamW, have shown to offer substantial benefits for models like BERT and ALBERT, as evidenced by improved F1 scores, indicating a stronger model performance due to the enhanced stability larger batches provide [97]. This aligns with the observation that these models can better utilize the stability and computational efficiency offered by larger batch sizes, optimizing the learning process more effectively. Conversely, models like DistilBERT and Electra do not show a marked preference for batch size adjustments, suggesting an intrinsic efficiency that may stem from their design and pre-training strategies. These findings imply that while BERT and ALBERT benefit significantly from larger batch sizes, the performance of DistilBERT and Electra is less contingent on this hyperparameter, possibly due to differences in model architecture or optimization needs. Interestingly, the role of epsilon values in the optimization process does not exhibit a clear pattern across models, hinting that its impact may be overshadowed by the more pronounced effects of learning rates and batch sizes. This suggests a complex interplay of factors that influence model performance beyond simple hyperparameter adjustments.

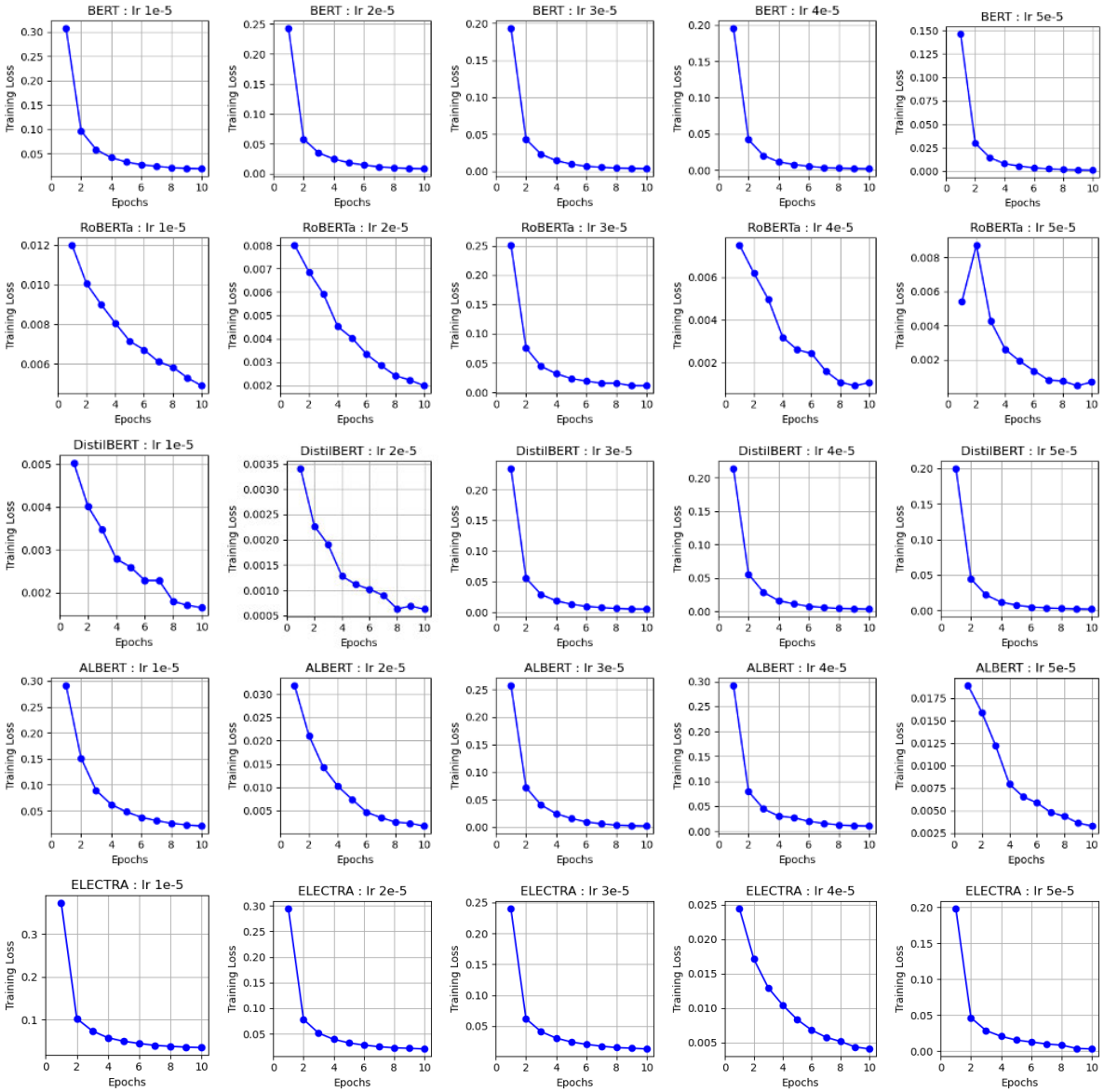


FIGURE 8. Training curves of different transformer models in NER.

The research indicates that AdamW, particularly with larger batch sizes and lower learning rates, is more conducive to optimizing BERT and ALBERT models, highlighting the importance of fine-tuning optimization strategies for peak performance. This preference suggests a reliance on precise optimization techniques to achieve optimal results. In contrast, RoBERTa's performance, while also benefiting from AdamW, suggests a degree of inherent robustness within its architecture, possibly making it less sensitive to specific optimizer configurations. The differential impact of optimization strategies on these models underscores the importance of tailored approaches in leveraging the full capabilities of each model, with specific attention to batch size and optimizer selection as key factors in maximizing performance outcomes [98].

In summary, while AdamW generally provides a performance edge, the extent of its benefits varies by model, with BERT and ALBERT showing the greatest improvements, RoBERTa demonstrating moderate sensitivity, and DistilBERT and Electra indicating a more optimizer-agnostic behavior. This reflects the complex interplay between model architecture and optimization techniques, underscoring the necessity for model-specific hyper-parameter tuning.

G. MODELS' TRAINING EVALUATION

Figure 8 provides illustrates the training curves for various NER models, with each subplot representing a different model or a variant of a model under different learning rates.

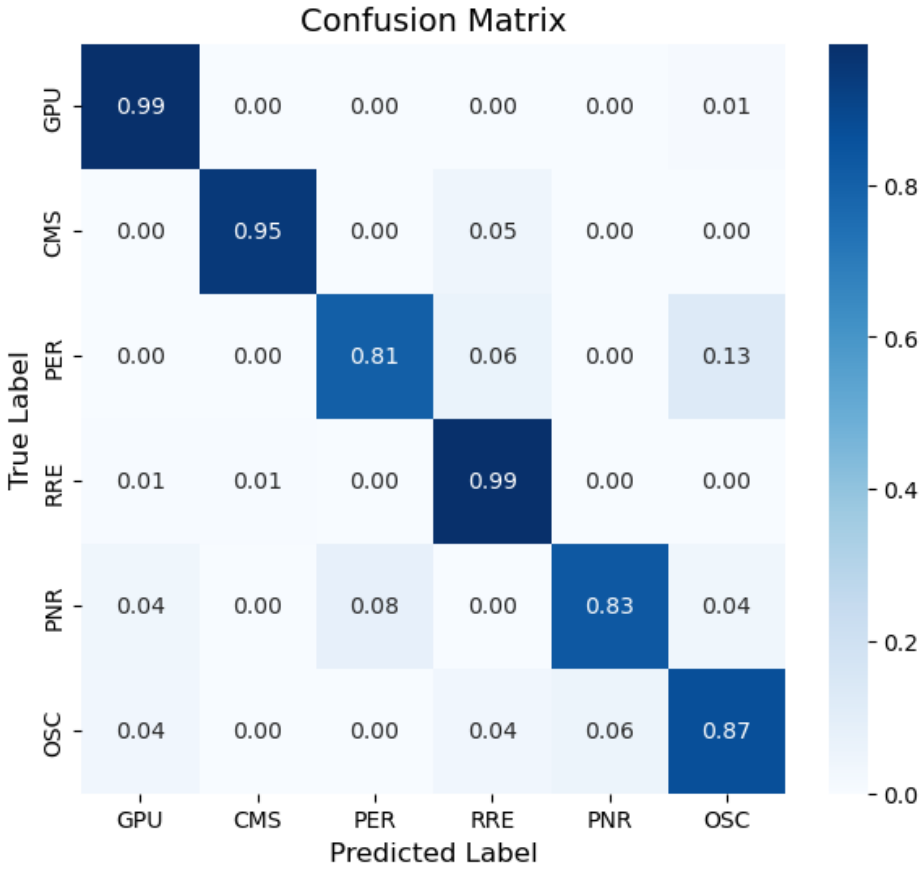


FIGURE 9. Heat map of confusion matrix of BERT model.

The models showcased are BERT, RoBERTa, DistilBERT, ALBERT, and ELECTRA, with learning rates ranging from 1e-5 to 5e-5. These models are evidently being applied to the domain of construction supply chain risk management, focusing on news article datasets.

Upon closer examination of the training curves, it's evident that the increase in learning rate from 1e-5 to 5e-5 significantly impacts the speed at which models like BERT and ELECTRA reduce their training loss. This phenomenon, marked by a steeper initial descent in loss curves, suggests an enhanced learning efficiency within the specified learning rate range. However, the increase to the higher learning rate of 5e-5 introduces a notable volatility in the loss reduction for BERT and ELECTRA models. This volatility could imply that these models are either reaching the saturation point of their learning capacity for the given dataset or are facing challenges in efficiently navigating the error landscape at such an accelerated rate of learning. This observation is critical, as it highlights the delicate balance between speeding up the training process and ensuring stable learning progress without compromising the model's ability to generalize from the training data.

On the contrary, models like RoBERTa and ALBERT exhibit a different behavior under the same conditions. They demonstrate smoother and more consistent training loss decreases across all examined learning rates, suggesting a superior ability to manage the complexities of the dataset without being overly sensitive to changes in the learning rate. This distinction in behavior can be attributed to the inherent architectural and pre-training differences among these models, which may influence their adaptability and resilience to learning rate adjustments. For instance, ALBERT's parameter-reduction techniques designed to lower memory consumption and increase training speed might contribute to its smoother learning curve. Similarly, RoBERTa's optimized pre-training approach could play a role in its ability to maintain stable loss reduction across varying learning rates. These observations underscore the importance of carefully selecting and tuning the learning rate, as it has a profound impact on the training dynamics and ultimate performance of language models. Understanding these nuances can guide more efficient and effective training strategies, tailoring the learning rate to match the specific characteristics and capacities of different models.

Comparing the models directly, RoBERTa and ALBERT appear to achieve lower training losses at the end of 10 epochs for all learning rates when compared to BERT, DistilBERT, and ELECTRA. This might indicate their greater

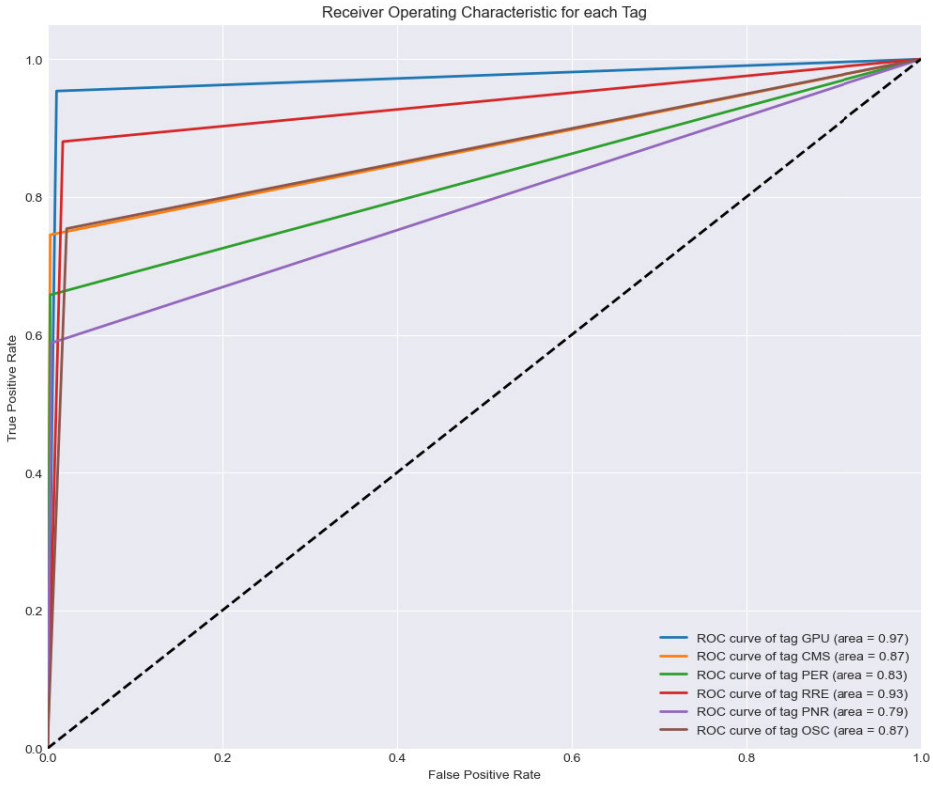


FIGURE 10. Receiver Operating Characteristic (ROC) curve of BERT model.

efficacy in capturing the nuances of named entity recognition within the specialized context of construction supply chain risk management. DistilBERT, while showing a favourable decline in training loss, does not reach as low a value as RoBERTa or ALBERT, which may be due to its distilled nature and thus a reduced capacity to model complex patterns in the data. ELECTRA, albeit starting with a high training loss especially at lower learning rates, demonstrates a significant reduction as training progresses, suggesting its potential effectiveness given sufficient training time.

H. ANALYSIS OF BERT MODEL'S PERFORMANCE

The analysis of the confusion matrix for the BERT model in NER task, as depicted in Figure 9 and trained using the optimal combination of hyper-parameters as discussed in Section IV-F, reveals crucial insights into the model's performance and areas for potential enhancement. Notably, a significant strength of the model is its ability to accurately predict certain labels, particularly RRE and GPU. The high accuracy in classifying RRE, which pertains to risks identified by the risk taxonomy related to the construction supply chain, demonstrates the model's adeptness in learning and distinguishing the unique characteristics of this specific entity class. This proficiency is indicative of the model's capacity to discern and accurately identify entities that are critical in the context of construction supply chain management. Such an ability is instrumental in ensuring the effectiveness of risk management processes within this domain, as it facilitates the accurate and timely identification of potential risks, thereby enabling more informed and strategic decision-making. The model's success in precisely classifying RRE entities can be attributed to its sophisticated learning algorithms and the fine-tuning of hyper-parameters, which collectively contribute to its nuanced understanding and recognition of these complex entity types.

Another positive point is the potential for the model to improve in areas where confusion is currently observed, such as the misclassification between GPU and OSC, and the misidentification of PER as RRE. These instances of confusion, while highlighting current limitations, also present opportunities for enhancement. As noted in existing research, confusion in NER tasks often stems from word ambiguity and the model's challenge in context-based disambiguation [99]. This suggests that the model has a foundational ability to process and classify entities but requires further refinement in handling contextual nuances.

The application of advanced techniques such as contrastive learning can significantly augment the model's capability to discern subtle differences between similar entities. Moreover, incorporating knowledge-guided approaches and adopting a question-answering framework in the training process, as found in research [100], can further enhance the model's performance. These methods provide a more comprehensive linguistic context, enabling the model to better understand

and differentiate complex entity structures. By integrating these improvements, the model's ability to accurately classify entities, particularly those currently prone to confusion, can be substantially enhanced, thus leveraging its existing strengths while addressing its limitations.

Figure 10 not only illustrates the efficacy of the BERT model in NER tasks but also provides a comparative analysis highlighting its versatility and precision in identifying various entity types. The Receiver Operating Characteristic (ROC) curves, differentiated by color for each tag, serve as a visual testament to the model's capability to discern between relevant and irrelevant entities with remarkable accuracy. The Area Under the Curve (AUC) scores, which range from 0.79 to 0.97, underscore a significant degree of model proficiency, with the ''GPU'' tag achieving an AUC of 0.97, nearly perfect in its predictive capability. This exceptional performance on the ''GPU'' tag indicates BERT's adeptness at distinguishing entities with high specificity, minimizing false positive rates, and effectively handling data imbalances.

Furthermore, the model's performance across a variety of other tags such as ''CMS'', ''PER'', ''RRE'', and ''OSC'' with AUC scores consistently above 0.8 reveals its broad applicability and robustness in processing and classifying diverse types of entities. This versatility is particularly important in complex NER tasks where the context and subtle nuances of language play a critical role in accurate entity recognition. BERT's architecture, leveraging pre-trained language models on vast corpora, enables it to capture and utilize these nuances, setting a benchmark in the field. The comparative analysis further highlights the model's strengths and areas of excellence, showcasing its leading-edge performance in the landscape of natural language processing tools. Such insights affirm the advanced capabilities of BERT in handling NER tasks, illustrating its pivotal role in advancing text analysis and processing technologies.

V. CONCLUSION AND FUTURE DIRECTIONS

This study has demonstrated the effectiveness of various transformer-based models in NER within the Australian construction supply chain risk management context, specifically using news articles. Models like BERT, RoBERTa, DistilBERT, ALBERT, and ELECTRA were evaluated, highlighting their respective strengths in processing and analyzing textual data for risk identification and management. A limitation of this study is the exclusion of the T5 and GPT-3 models from the grid search analysis.

Furthermore, through the integration of advanced NER technologies, such as BERT, RoBERTa, and ELECTRA, the Australian construction sector enhances its ability to navigate international market dynamics and geopolitical shifts, thereby enabling more resilient and responsive supply chain operations through timely risk identification and proactive management strategies.

A. MODEL PERFORMANCE

The comparative analysis of different transformer models revealed varying levels of efficacy in NER tasks. Models like BERT and RoBERTa showed robust performance, particularly in terms of precision and recall, indicating their suitability for extracting relevant entities from complex textual data. These insights are crucial for advancing the field of NER and its application in construction supply chain risk management.

B. PROJECT MANAGEMENT PERFORMANCE

Sophisticated transformer models like BERT, RoBERTa, and ELECTRA have revolutionized project management, particularly in the construction industry. Their ability to analyze vast amounts of text data, including global news trends, allows project managers to detect early warning signs of potential disruptions in the supply chain. This capability is crucial in areas like the Australian construction sector, where international market dynamics and geopolitical shifts significantly impact operations. By leveraging NER technologies, project managers gain a nuanced understanding of the supply chain environment, enhancing their ability to navigate risk factors efficiently, and ensuring projects stay on course amidst global uncertainties. These tools not only provide detailed risk profiles but also make project planning more agile, helping managers to stay aligned with timelines and budgets.

The adaptability of transformer models to specific domains, as shown in various studies, allows for fine-tuning to recognize unique terminologies and nuances in the construction industry. Integrating these advanced NER systems into analytical frameworks significantly enhances risk assessment and management strategies. Practitioners can sift through news data to extract vital information such as disruptions, market changes, or regulatory updates. This enhanced entity recognition capability means even subtle references or complex entity relationships in news articles are accurately identified, offering a comprehensive view of potential risks and opportunities. Consequently, this leads to more resilient and responsive supply chain operations, empowering decision-makers with timely insights and fostering a proactive approach to risk management in the construction supply chain.

In other words, the findings underscore the importance of NER in enhancing supply chain resilience in the construction industry. As highlighted by [101], recognizing company names from textual data is challenging due to the diverse ways a company can be referenced. NER systems that can accurately identify these entities are crucial in risk management, especially for non-exchange-listed entities where obtaining timely information is challenging. The use of NER in constructing company-relationship graphs is particularly beneficial in risk management within institutions, allowing for better understanding and mitigation of risks in the supply chain.

Secondly, the deployment of NER systems contributes significantly to the proactive management of supply chain risks. [102] emphasize the significance of monitoring industry-relevant events for supply chain management. NER facilitates the extraction of specific events and entities from high-volume, heterogeneous text streams, such as traffic reports, tweets, and news articles, which are crucial for anticipating and managing potential disruptions in the construction supply chain. By enabling the extraction of fine-grained entities and relationships, NER systems enhance the responsiveness and adaptability of supply chain risk management strategies.

All in all, the deployment of transformer models with enhanced entity recognition capability is instrumental in driving construction supply chain responsiveness, as evidenced by recent studies [103], [104], [105], [106]. This technological advancement is pivotal in navigating the complexities of supply chain management, enhancing risk identification and mitigation strategies, ensuring sustained operational efficiency amid disruptions.

C. FUTURE DIRECTIONS

In the exploration of alternative transformer models, further studies can consider trying out XLNet and DeBERTa. XLNet's strength lies in its ability to capture long-range dependencies and handle previously unseen entities, making it highly effective for NER tasks in varied domains, including complex and technical ones like construction [107], [108], [109]. This capability is particularly useful in construction project management, where the ability to accurately identify and classify entities from project documentation can enhance risk management and sequence planning.

DeBERTa, on the other hand, improves upon BERT and XLNet by incorporating disentangled attention mechanisms, enabling more precise and contextually aware entity recognition [110]. This enhanced attention to contextual detail is crucial in the construction industry, where documents often contain a mix of technical terms, project-specific jargon, and standard language. DeBERTa's advanced feature extraction and contextual understanding can lead to more accurate identification of risks and entities, thus contributing to more efficient project management.

- To further improve the performance of NER models in construction risk management, a more detailed strategy for tuning hyperparameters could be employed. This involves expanding the scope of the grid search to consider a wider range of parameters for the Adam and AdamW optimizers. For example, different weight decay rates or learning rate schedules could be explored.
- The integration of NER capabilities into project management software could greatly improve the risk identification process. This integration would provide project managers with real-time alerts and suggestions, leveraging the latest news and market trends. As a result, project management becomes more proactive and adaptive.
- Sentiment analysis is a valuable tool for risk assessment. By combining NER with sentiment analysis, we can gain a better understanding of the potential impact of identified risks. By assessing the sentiment of news articles and reports, we can prioritise risks based on their urgency.
- To enhance the performance of transformer models such as BERT, T5, GPT-3, RoBERTa, DistilBERT, and ELECTRA in NER, future studies can adopt a variety of innovative approaches. Ensemble methods that combine multiple transformer-based models can demonstrate significant improvements in performance, especially in handling complex, cross-domain texts. Integrating language model pretraining with NER fine-tuning would be another effective strategy, particularly in enhancing cross-domain and cross-lingual generalization abilities of these models. Additionally, the adoption of multi-task learning approaches, where auxiliary tasks like entity boundary prediction and entity type prediction are integrated into the initial layers of the transformer, can show promise in better leveraging the lower layers for more robust character representation. Furthermore, the development of noise reduction models, particularly those utilizing XLNet encoding and CRF decoding, can effectively addresse the challenges of noise interference in both the pre-training and fine-tuning stages.

DECLARATION OF COMPETING INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

DATA AVAILABILITY

Data will be made available on request.

REFERENCES

[1] J. Hirschberg and C. D. Manning, "Advances in natural language processing," *Science*, vol. 349, pp. 261–266, Jul. 2015.
[2] P. M. Nadkarni, L. Ohno-Machado, and W. W. Chapman, "Natural language processing: An introduction," *J. Amer. Med. Inform. Assoc.*, vol. 18, no. 5, pp. 544–551, 2011.
[3] M. Zampieri, P. Nakov, and Y. Scherrer, "Natural language processing for similar languages, varieties, and dialects: A survey," *Natural Lang. Eng.*, vol. 26, no. 6, pp. 595–612, Nov. 2020.
[4] A. N. M. Fahim Faisal, Md. A. Rahman, and T. Farah, "A rule-based Bengali grammar checker," in *Proc. 5th World Conf. Smart Trends Syst. Secur. Sustainability*, Jul. 2021, pp. 113–117.
[5] D. Newman-Griffis, J. F. Lehman, C. Rosé, and H. Hochheiser, "Translational NLP: A new paradigm and general principles for natural language processing research," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2021, p. 4125.
[6] D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural language processing: State of the art, current trends and challenges," *Multimedia Tools Appl.*, vol. 82, no. 3, pp. 3713–3744, Jan. 2023.
[7] E. Ghazizadeh and P. Zhu, "A systematic literature review of natural language processing: Current state, challenges and risks," in *Proc. Future Technol. Conf.* Cham, Switzerland: Springer, 2020, pp. 634–647.
[8] K. Chowdhary and K. Chowdhary, "Natural language processing," in *Fundamentals of Artificial Intelligence*. New Delhi: Springer, 2020, pp. 603–649.

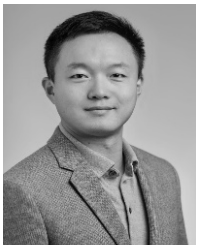
[9] C. Biemann, *Theory and Applications of Natural Language Processing*. New Delhi: Springer, 2010.
[10] S. Francis, J. V. Landeghem, and M.-F. Moens, "Transfer learning for named entity recognition in financial and biomedical documents," *Information*, vol. 10, no. 8, p. 248, Jul. 2019.
[11] M. V. Koroteev, "BERT: A review of applications in natural language processing and understanding," 2021, *arXiv:2103.11943*.
[12] S. O. Abioye, L. O. Oyedele, L. Akanbi, A. Ajayi, J. M. D. Delgado, M. Bilal, O. O. Akinade, and A. Ahmed, "Artificial intelligence in the construction industry: A review of present status, opportunities and future challenges," *J. Building Eng.*, vol. 44, Dec. 2021, Art. no. 103299.
[13] J. P. C. Chiu and E. Nichols, "Named entity recognition with bidirectional LSTM-CNNs," *Trans. Assoc. Comput. Linguistics*, vol. 4, pp. 357–370, Dec. 2016.
[14] S. Wang, R. Xu, B. Liu, L. Gui, and Y. Zhou, "Financial named entity recognition based on conditional random fields and information entropy," in *Proc. Int. Conf. Mach. Learn. Cybern.*, vol. 2, Jul. 2014, pp. 838–843.
[15] M. Miwa and M. Bansal, "End-to-end relation extraction using LSTMs on sequences and tree structures," 2016, *arXiv:1601.00770*.
[16] M. E. Peters, M. Neumann, L. Zettlemoyer, and W.-T. Yih, "Dissecting contextual word embeddings: Architecture and representation," 2018, *arXiv:1808.08949*.
[17] T. B. Brown et al., "Language models are few-shot learners," in *Proc. NIPS*, 2020, pp. 1877–1901.
[18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.
[19] C. Sutton, "An introduction to conditional random fields," *Found. Trends Mach. Learn.*, vol. 4, no. 4, pp. 267–373, 2012.
[20] P. X. W. Zou and P. Couani, "Managing risks in green building supply chain," *Architectural Eng. Des. Manage.*, vol. 8, no. 2, pp. 143–158, May 2012.
[21] T.-K. Wang, Q. Zhang, H.-Y. Chong, and X. Wang, "Integrated supplier selection framework in a resilient construction supply chain: An approach via analytic hierarchy process (AHP) and grey relational analysis (GRA)," *Sustainability*, vol. 9, no. 2, p. 289, Feb. 2017.
[22] P. X. W. Zou, D. McGeorge, and S. Ng, "Small and medium-sized enterprises' perspectives towards construction supply chain management and e-commerce," *Int. J. Construct. Manage.*, vol. 5, no. 1, pp. 1–19, Jan. 2005.
[23] C. Berragan, A. Singleton, A. Calafiore, and J. Morley, "Transformer based named entity recognition for place name extraction from unstructured text," *Int. J. Geographical Inf. Sci.*, vol. 37, no. 4, pp. 747–766, Apr. 2023.
[24] C.-M. Tsai, "Stylometric fake news detection based on natural language processing using named entity recognition: In-domain and cross-domain analysis," *Electronics*, vol. 12, no. 17, p. 3676, Aug. 2023.
[25] Y. Bengio, R. Ducharme, and P. Vincent, "A neural probabilistic language model," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 13, 2000, pp. 1–12.
[26] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, *arXiv:1301.3781*.
[27] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 26, 2013, pp. 1–13.
[28] S. J. Johnson, M. R. Murty, and I. Navakanth, "A detailed review on word embedding techniques with emphasis on word2vec," *Multimedia Tools Appl.*, vol. 2023, pp. 1–29, Oct. 2023.
[29] F. Almeida and G. Xexéo, "Word embeddings: A survey," 2019, *arXiv:1901.09069*.
[30] T.-H. Yang, M. Pleva, D. Hládek, and M.-H. Su, "BERT-based Chinese medicine named entity recognition model applied to medication reminder dialogue system," in *Proc. 13th Int. Symp. Chin. Spoken Lang. Process. (ISCSLP)*, Dec. 2022, pp. 374–378.
[31] C. Yang, L. Sheng, Z. Wei, and W. Wang, "Chinese named entity recognition of epidemiological investigation of information on COVID-19 based on BERT," *IEEE Access*, vol. 10, pp. 104156–104168, 2022.
[32] D. Chen, C. Liu, and Z. Zhao, "Named entity recognition service of BERT-transformer-CRF based on multi-feature fusion for chronic disease management," in *Proc. Int. Conf. Service Sci.* Cham, Switzerland: Springer, 2023, pp. 166–178.
[33] Y. Yu, Y. Wang, J. Mu, W. Li, S. Jiao, Z. Wang, P. Lv, and Y. Zhu, "Chinese mineral named entity recognition based on BERT model," *Expert Syst. Appl.*, vol. 206, Nov. 2022, Art. no. 117727.
[34] X. Tang, Y. Huang, M. Xia, and C. Long, "A multi-task BERT-BiLSTM-AM-CRF strategy for Chinese named entity recognition," *Neural Process. Lett.*, vol. 55, no. 2, pp. 1209–1229, Apr. 2023.
[35] S. Gorla, S. S. Tangeda, L. B. M. Neti, and A. Malapati, "Telugu named entity recognition using bert," *Int. J. Data Sci. Anal.*, vol. 14, no. 2, pp. 127–140, Aug. 2022.
[36] M. Jarrar, M. Khalilia, and S. Ghanem, "Wojood: Nested Arabic named entity corpus and recognition using BERT," in *Proc. 13th Lang. Resour. Eval. Conf.*, 2022, pp. 3626–3636.
[37] Z. Lun and Z. Hui, "Research on agricultural named entity recognition based on pre train BERT," *Academic J. Eng. Technol. Sci.*, vol. 5, no. 4, pp. 34–42, 2022.
[38] C. V. Ndukwe, J. Liu, and T. K. Chan, "Impact of COVID-19 on the China–Australia construction supply chain," in *Proc. 25th Int. Symp. Advancement Construct. Manage. Real Estate*. Cham, Switzerland: Springer, 2021, pp. 1275–1291.
[39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.
[40] C. Huang, Y. Wang, Y. Yu, Y. Hao, Y. Liu, and X. Zhao, "Chinese named entity recognition of geological news based on BERT model," *Appl. Sci.*, vol. 12, no. 15, p. 7708, Jul. 2022.
[41] Q. Qiu, Z. Xie, L. Wu, and L. Tao, "Automatic spatiotemporal and semantic information extraction from unstructured geoscience reports using text mining techniques," *Earth Sci. Informat.*, vol. 13, no. 4, pp. 1393–1410, Dec. 2020.
[42] X. Lv, Z. Xie, D. Xu, X. Jin, K. Ma, L. Tao, Q. Qiu, and Y. Pan, "Chinese named entity recognition in the geoscience domain based on BERT," *Earth Space Sci.*, vol. 9, no. 3, Mar. 2022, Art. no. e2021EA002166.
[43] R. Fan, L. Wang, J. Yan, W. Song, Y. Zhu, and X. Chen, "Deep learning-based named entity recognition and knowledge graph construction for geological hazards," *ISPRS Int. J. Geo-Inf.*, vol. 9, no. 1, p. 15, Dec. 2019.
[44] X. Su, Z. Hong, Q. Zhang, C. Xue, and X. Li, "Named entity recognition for Chinese construction documents," in *Proc. Int. Symp. Advancement Construct. Manage. Real Estate*. Cham, Switzerland: Springer, 2019, pp. 839–850.
[45] P. Schönfelder and M. König, "Deep learning-based entity recognition in construction regulatory documents," in *Proc. 38th Int. Symp. Autom. Robot. Construct. (ISARC)*, Nov. 2021, pp. 387–394.
[46] K. Jeon, G. Lee, S. Yang, and H. D. Jeong, "Named entity recognition of building construction defect information from text with linguistic noise," *Autom. Construct.*, vol. 143, Nov. 2022, Art. no. 104543.
[47] X. Wu, T. Zhang, S. Yuan, and Y. Yan, "One improved model of named entity recognition by combining BERT and BiLSTM-CNN for domain of Chinese railway construction," in *Proc. 7th Int. Conf. Intell. Comput. Signal Process. (ICSP)*, Apr. 2022, pp. 728–732.
[48] Q. Zhang, C. Xue, X. Su, P. Zhou, X. Wang, and J. Zhang, "Named entity recognition for Chinese construction documents based on conditional random field," *Frontiers Eng. Manage.*, vol. 10, no. 2, pp. 237–249, Jun. 2023.
[49] X. Wang and N. El-Gohary, "Deep learning–based named entity recognition and resolution of referential ambiguities for enhanced information extraction from construction safety regulations," *J. Comput. Civil Eng.*, vol. 37, no. 5, Sep. 2023, Art. no. 04023023.
[50] K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, and Y. Wang, "Transformer in transformer," in *Proc. NIPS*, 2021, pp. 15908–15919.
[51] A. Chernyavskiy, D. Ilvovsky, and P. Nakov, "Transformers: 'The end of history' for natural language processing?" in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*, Bilbao, Spain. Cham, Switzerland: Springer, 2021, pp. 677–693.
[52] K. Huangliang, X. Li, T. Yin, B. Peng, and H. Zhang, "Self-adapted positional encoding in the transformer encoder for named entity recognition," in *Proc. Int. Conf. Artif. Neural Netw.* Cham, Switzerland: Springer, 2023, pp. 538–549.
[53] B. Ghojogh and A. Ghodsi, "Attention mechanism, transformers, BERT, and GPT: Tutorial and survey," Center Open Sci., Charlottesville, VA, USA, Tech. Rep., 2020, doi: 10.31219/osf.io/m6gcn.
[54] N. Mohammadi Foumani, C. Wei Tan, G. I. Webb, and M. Salehi, "Improving position encoding of transformers for multivariate time series classification," 2023, *arXiv:2305.16642*.
[55] S. N. M. Foumani and A. Nickabadi, "A probabilistic topic model using deep visual word representation for simultaneous image classification and annotation," *J. Vis. Commun. Image Represent.*, vol. 59, pp. 195–203, Feb. 2019.

[56] N. Sabharwal, A. Agrawal, N. Sabharwal, and A. Agrawal, "Bert algorithms explained," in *Hands-on Question Answering Systems With BERT: Applications in Neural Networks and Natural Language Processing*. New York, NY, USA: Apress, 2021, pp. 65–95.
[57] B. Wang, L. Shang, C. Lioma, X. Jiang, H. Yang, Q. Liu, and J. G. Simonsen, "On position embeddings in BERT," in *Proc. Int. Conf. Learn. Represent.*, 2020.
[58] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.
[59] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," 2019, *arXiv:1910.01108*.
[60] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A lite BERT for self-supervised learning of language representations," pp. 1–5, 2019, *arXiv:1909.11942*.
[61] 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*.
[62] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, no. 1, pp. 5485–5551, 2020.
[63] S. Wang, X. Sun, X. Li, R. Ouyang, F. Wu, T. Zhang, J. Li, and G. Wang, "GPT-NER: Named entity recognition via large language models," 2023, *arXiv:2304.10428*.
[64] M. Zhang and J. Li, "A commentary of GPT-3 in MIT technology review 2021," *Fundam. Res.*, vol. 1, no. 6, pp. 831–833, Nov. 2021.
[65] A. Coburn, D. Ralph, M. Tuveson, S. Ruffle, and G. Bowman, "A taxonomy of threats for macro-catastrophe risk management," Judge Bus. School, Cambridge Centre Risk Stud. Work. Paper Series, Work. Paper 201307.20, 2013, pp. 20–24.
[66] D. Ortiz, D. Myers, E. Walls, and M.-E. Diaz, "Where do we stand with newspaper data?" *Mobilization, Int. Quart.*, vol. 10, no. 3, pp. 397–419, Oct. 2005.
[67] C. Feng, M. Khan, A. U. Rahman, and A. Ahmad, "News recommendation systems—Accomplishments, challenges & future directions," *IEEE Access*, vol. 8, pp. 16702–16725, 2020.
[68] E. F. T. K. Sang and F. De Meulder, "Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition," 2003, *arXiv:cs/0306050*.
[69] L. Volkova and V. Bocharov, "An approach to inter-annotator agreement evaluation for the named entities annotation task at OpenCorpora," in *Artificial Intelligence and Natural Language* (Communications in Computer and Information Science). New Delhi: Springer, 2019.
[70] S. Moon, G. Lee, S. Chi, and H. Oh, "Automated construction specification review with named entity recognition using natural language processing," *J. Construct. Eng. Manage.*, vol. 147, no. 1, Jan. 2021, Art. no. 04020147.
[71] W. Ding, M. Abdel-Basset, H. Hawash, and A. M. Ali, "Explainability of artificial intelligence methods, applications and challenges: A comprehensive survey," *Inf. Sci.*, vol. 615, pp. 238–292, 2022.
[72] K. Clark and T. Luong. (2020). *More Efficient NLP Model Pre-Training With Electra*. [Online]. Available: https://ai.googleblog.com/2020/03/more-efficient-nlp-model-pre-training.html
[73] C. Lothritz, K. Allix, L. Veiber, T. F. Bissyandé, and J. Klein, "Evaluating pretrained transformer-based models on the task of fine-grained named entity recognition," in *Proc. 28th Int. Conf. Comput. Linguistics*, 2020, pp. 3750–3760.
[74] M. Abadeer, "Assessment of DistilBERT performance on named entity recognition task for the detection of protected health information and medical concepts," in *Proc. 3rd Clin. Natural Lang. Process. Workshop*, 2020, pp. 158–167.
[75] B. S. Jati, S. Widyawan, and S. T. M. N. Rizal, "Multilingual named entity recognition model for Indonesian health insurance question answering system," in *Proc. 3rd Int. Conf. Inf. Commun. Technol. (ICOIACT)*, Nov. 2020, pp. 180–184.
[76] A. Sartipi and A. Fatemi, "Exploring the potential of machine translation for generating named entity datasets: A case study between Persian and English," 2023, *arXiv:2302.09611*.
[77] S. Singh, P. Jawale, and U. Tiwary, "Silpa_nlp at SemEval-2022 tasks 11: Transformer based NER models for Hindi and Bangla languages," in *Proc. 16th Int. Workshop Semantic Eval.*, 2022, pp. 1536–1542.
[78] N. Minh Lai, "LMN at SemEval-2022 task 11: A transformer-based system for English named entity recognition," 2022, *arXiv:2203.03546*.
[79] E. Tavan and M. Najafi, "MarSan at SemEval-2022 task 11: Multilingual complex named entity recognition using T5 and transformer encoder," in *Proc. 16th Int. Workshop Semantic Eval.*, 2022, pp. 1639–1647.
[80] I. Ashrafi, M. Mohammad, A. S. Mauree, G. M. A. Nijhum, R. Karim, N. Mohammed, and S. Momen, "Banner: A cost-sensitive contextualized model for Bangla named entity recognition," *IEEE Access*, vol. 8, pp. 58206–58226, 2020.
[81] H. Patel, "BioNerFlair: Biomedical named entity recognition using flair embedding and sequence tagger," 2020, *arXiv:2011.01504*.
[82] E. Schneider, R. M. Rivera-Zavala, P. Martinez, C. Moro, and E. Paraiso, "UC3M-PUCPR at SemEval-2022 task 11: An ensemble method of transformer-based models for complex named entity recognition," in *Proc. 16th Int. Workshop Semantic Eval.*, 2022, pp. 1448–1456.
[83] X. Han, Q. Yue, J. Chu, Z. Han, Y. Shi, and C. Wang, "Multi-feature fusion transformer for Chinese named entity recognition," in *Proc. 41st Chin. Control Conf. (CCC)*, Jul. 2022, pp. 4227–4232.
[84] J. Li, Q. Wei, O. Ghiasvand, M. Chen, V. Lobanov, C. Weng, and H. Xu, "A comparative study of pre-trained language models for named entity recognition in clinical trial eligibility criteria from multiple corpora," *BMC Med. Informat. Decis. Making*, vol. 22, no. S3, pp. 1–10, Sep. 2022.
[85] F. Deng, D. Zhang, and J. Peng, "Biological named entity recognition and role labeling via deep multi-task learning," in *Proc. 13th Int. Conf. Mach. Learn. Comput.*, Feb. 2021, pp. 450–455.
[86] H. Tan, Z. Yang, J. Ning, Z. Ding, and Q. Liu, "Chinese medical named entity recognition based on Chinese character radical features and pre-trained language models," in *Proc. Int. Conf. Asian Lang. Process. (IALP)*, Dec. 2021, pp. 121–124.
[87] A. Youssef, M. Elattar, and S. R. El-Beltagy, "A multi-embeddings approach coupled with deep learning for Arabic named entity recognition," in *Proc. 2nd Novel Intell. Lead. Emerg. Sci. Conf. (NILES)*, Oct. 2020, pp. 456–460.
[88] A. A. Choure, R. B. Adhao, and V. K. Pachghare, "NER in Hindi language using transformer model: XLM-RoBERTa," in *Proc. IEEE Int. Conf. Blockchain Distrib. Syst. Secur. (ICBDS)*, Sep. 2022, pp. 1–5.
[89] D. Cortiz, "Exploring transformers models for emotion recognition: A comparision of BERT, DistilBERT, RoBERTa, XLNET and ELECTRA," in *Proc. 3rd Int. Conf. Control, Robot. Intell. Syst.*, Aug. 2022, pp. 230–234.
[90] A. F. Adoma, N.-M. Henry, and W. Chen, "Comparative analyses of BERT, RoBERTa, DistilBERT, and XLNet for text-based emotion recognition," in *Proc. 17th Int. Comput. Conf. Wavelet Act. Media Technol. Inf. Process. (ICCWAMTIP)*, Dec. 2020, pp. 117–121.
[91] S. H. Oh, M. Kang, and Y. Lee, "Protected health information recognition by fine-tuning a pre-training transformer model," *Healthcare Informat. Res.*, vol. 28, no. 1, pp. 16–24, Jan. 2022.
[92] Y. Wu, J. Huang, C. Xu, H. Zheng, L. Zhang, and J. Wan, "Research on named entity recognition of electronic medical records based on RoBERTa and radical-level feature," *Wireless Commun. Mobile Comput.*, vol. 2021, pp. 1–10, Jun. 2021.
[93] C.-M. Huang, Y.-J. Lee, D. K. J. Lin, and S.-Y. Huang, "Model selection for support vector machines via uniform design," *Comput. Statist. Data Anal.*, vol. 52, no. 1, pp. 335–346, Sep. 2007.
[94] A. John Quijano, S. Nguyen, and J. Ordonez, "Grid search hyperparameter benchmarking of BERT, ALBERT, and LongFormer on DuoRC," 2021, *arXiv:2101.06326*.
[95] L. Bottou, F. E. Curtis, and J. Nocedal, "Optimization methods for large-scale machine learning," *SIAM Rev.*, vol. 60, no. 2, pp. 223–311, Jan. 2018.
[96] G. Yenduri, R. M. C. Selvi, G. Srivastava, P. K. R. Maddikunta, D. Raj, R. H. Jhaveri, W. Wang, A. V. Vasilakos, and T. Reddy Gadekallu, "Generative pre-trained transformer: A comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions," 2023, *arXiv:2305.10435*.
[97] S. Zheng, H. Lin, S. Zha, and M. Li, "Accelerated large batch optimization of BERT pretraining in 54 minutes," 2020, *arXiv:2006.13484*.
[98] Z. Nado, J. M. Gilmer, C. J. Shallue, R. Anil, and G. E. Dahl, "A large batch optimizer reality check: Traditional, generic optimizers suffice across batch sizes," 2021, *arXiv:2102.06356*.
[99] R. Zhou, Q. Hu, J. Wan, J. Zhang, Q. Liu, T. Hu, and J. Li, "WCL-BBCD: A contrastive learning and knowledge graph approach to named entity recognition," 2022, *arXiv:2203.06925*.

[100] P. Banerjee, K. K. Pal, M. Devarakonda, and C. Baral, "Biomedical named entity recognition via knowledge guidance and question answering," *ACM Trans. Comput. Healthcare*, vol. 2, no. 4, pp. 1–24, Oct. 2021.
[101] M. Loster, Z. Zuo, F. Naumann, O. Maspfuhl, and D. Thomas, "Improving company recognition from unstructured text by using dictionaries," in *Proc. EDBT*, 2017, pp. 610–619.
[102] M. Schiersch, V. Mironova, M. Schmitt, P. Thomas, A. Gabryszak, and L. Hennig, "A German corpus for fine-grained named entity recognition and relation extraction of traffic and industry events," 2020, *arXiv:2004.03283*.
[103] T. O. Osunsanmi, C. O. Aigbavboa, W. D. D. Thwala, and R. Molusiwa, "Modelling construction 4.0 as a vaccine for ensuring construction supply chain resilience amid COVID-19 pandemic," *J. Eng., Des. Technol.*, vol. 20, no. 1, pp. 132–158, Jan. 2022.
[104] M. O. Gani, T. Yoshi, and M. S. Rahman, "Optimizing firm's supply chain resilience in data-driven business environment," *J. Global Oper. Strategic Sourcing*, vol. 16, no. 2, pp. 258–281, Apr. 2023.
[105] M. Z. Alvarenga, M. P. V. D. Oliveira, and T. A. G. F. D. Oliveira, "The impact of using digital technologies on supply chain resilience and robustness: The role of memory under the COVID-19 outbreak," *Supply Chain Manag., Int. J.*, vol. 28, no. 5, pp. 825–842, Jul. 2023.
[106] Z.-P. Li, H.-T. Ceong, and S.-J. Lee, "The effect of blockchain operation capabilities on competitive performance in supply chain management," *Sustainability*, vol. 13, no. 21, p. 12078, Nov. 2021.
[107] R. Yan, X. Jiang, and D. Dang, "Named entity recognition by using XLNet-BiLSTM-CRF," *Neural Process. Lett.*, vol. 53, no. 5, pp. 3339–3356, Oct. 2021.
[108] S. Feng, S. Min, and G. Lei, "Named entity recognition model of Chinese clinical electronic medical record based on XLNet-BiLSTM," *Res. Square*, early access, 2021, doi: 10.21203/rs.3.rs-218833/v1.
[109] D. Yang, F. Wan, and Y. Zhang, "Named entity recognition in XLNet cyberspace security domain based on dictionary embedding," in *Proc. 4th Int. Conf. Adv. Comput. Technol., Inf. Sci. Commun. (CTISC)*, Apr. 2022, pp. 1–5.
[110] X.-D. Doan, "VTCC-NLP at NL4Opt competition subtask 1: An ensemble pre-trained language models for named entity recognition," 2022, *arXiv:2212.07219*.

MILAD BAGHALZADEH SHISHEHGARKHANEH is currently pursuing the Ph.D. degree in civil engineering with Monash University, Melbourne, Australia. Concurrently, he holds a lecturer position with Victoria University, Melbourne. His research is primarily focused on leveraging advanced artificial intelligence (AI) methodologies, notably transformer architectures, to address challenges in construction supply chain risk management. Additionally, he is exploring the integration of AI with blockchain technology to enhance resilience in construction projects. With over 28 published works, including journal articles and book chapters, he has made significant contributions to the field. His diverse research interests include construction supply chain management (CSCM), machine learning, natural language processing (NLP), blockchain technology, and building information modeling (BIM). Through his academic endeavors, he aims to drive innovation in construction management practices, ensuring projects are executed efficiently and resiliently amidst evolving industry challenges.

ROBERT C. MOEHLER is currently a Senior Lecturer in engineering project management with the Department of Infrastructure Engineering, Faculty of Engineering and Information Technology (FEIT), The University of Melbourne. His current projects include integration of a value approach through co-creation and co-design in the project-based industries; focusing on knowledge artifacts for community engagement; project process integration of information systems to enhance collaboration, and business model innovation through platform-supported project delivery. His work has appeared in IEEE TRANSACTIONS ON ENGINEERING MANAGEMENT, *Journal of Cleaner Production*, *International Journal of Construction Management*, *Project Management Journal*, and *Reliability Engineering and System Safety*.



YIHAI FANG received the Ph.D. degree in civil engineering from the Georgia Institute of Technology. He is currently a Senior Lecturer with the Department of Civil Engineering, Monash University. Following this, he was a Postdoctoral Associate with the University of Florida, before joining Monash University. His research interests include construction automation and informatics, construction robotics, and the digital twin for construction and built environments. Additionally, he has been leading various research projects exploring building productivity, the implementation of digital twin technology in facility management, and the evaluation of emerging technologies for remote inspections of building work. These endeavors underscore his commitment to fostering innovation and practical solutions within the construction industry.

AMER A. HIJAZI received the Bachelor of Engineering degree from Hashemite University, the Master of Science degree in BIM, design construction and operation from the University of the West of England, Bristol, and the Ph.D. degree from the Centre for Smart Modern Construction, Western Sydney University, Australia. He is currently an Assistant Professor with Al-Ahliyya Amman University, Jordan. He had a long-standing association with the industry and academia alike, with a decade of experience predominantly focused on management practices, with a strong emphasis on information technology and computer science in Australia, the U.K., and the MENA region. He completed his Ph.D. degree with a value of around 0.25 million, to develop a comprehensive framework that integrates blockchain capability suitable for the construction industry and its interrelation with BIM, to develop a BIM single source of truth (BIMSSoT) model. He aced the Blockchain Strategy Program offered by (Oxford University) and also the Digital Transformation: From AI and the IoT to cloud, blockchain, and cybersecurity programme by Massachusetts Institute of Technology, which helped him acquire the strategies he needs to respond to the latest developments in the construction sector. He has made significant contributions to several international research projects as part of the building 4.0 CRC in collaboration with Monash University, The University of Melbourne, and the Centre of Digital Built Britain, University of Cambridge.

HAMED ABOUTORAB received the Ph.D. degree from the University of New South Wales. He is currently a Cybersecurity Research Fellow with Charles Sturt University. His research interests include the integration of cybersecurity and artificial intelligence, with a current focus on security automation and orchestration. His research has been published in prestigious international journals, including IEEE TRANSACTIONS ON SERVICES COMPUTING, *Expert Systems with Applications*, *Journal of Network and Computer Applications*, and *Future Generation Computer Systems*.