ARTICLE OPEN

Check for updates

Searching for chromate replacements using natural language processing and machine learning algorithms

Shujing Zhao[1,2]✉ and Nick Birbilis[1,2,3]✉

The past few years have seen the application of machine learning utilised in the exploration of materials. As in many fields of research—the vast majority of knowledge is published as text, which poses challenges in either a consolidated or statistical analysis across studies and reports. To address this issue, the application of natural language processing (NLP) has been explored in several studies to date. In the present work, we have employed the Word2Vec model, previously explored by others, and the BERT model—applying them towards the search for chromate replacements in the field of corrosion protection. From a database of over 80 million records, a down-selection of 5990 papers focused on the topic of corrosion protection were examined using NLP. This study demonstrates it is possible to extract knowledge from the automated interpretation of the scientific literature and achieve expert human-level insights.

npj Materials Degradation (2023)7:2 ; https://doi.org/10.1038/s41529-022-00319-0

INTRODUCTION

The corrosion of metals and alloys remains a significant technological and financial issue—globally. Studies regarding the cost of corrosion, for example, those recently conducted in the USA[1], China[2], and Australia[3] reveal that corrosion costs amount to ~3% of GDP annually (which equates to global costs of >US$1 trillion per annum). As a result, methods for corrosion prevention remain critical. In terms of the corrosion protection of metals and alloys, for over half a century, the benchmark for exceptional performance from corrosion-inhibiting compounds has been demonstrated by hexavalent chromium (known as chromate)[4]. Chromate is a powerful inhibitor, as it can passivate many metals, including Zn, Al, Mg, etc. The mechanism of chromate protection involves the formation of a protective Cr (III) oxide layer on reactive metals from mobile and soluble Cr (VI) oxyanions that can migrate to 'active' (anode) sites[5]. Chromate serves as a corrosion inhibitor in aqueous solutions, but also as an additive to primers used to coat metals (such as steels and galvanised steels).

The International Agency for Research on Cancer (IARC) confirmed that hexavalent chromium (Cr (VI)) is a human carcinogen in 1990 based on independent studies around the world[6]. However, the corrosion inhibition performance of chromate-containing primers is appreciable, such that chromate-containing primers are the current industry benchmark in terms of performance (and additionally, consumer expectations of product performance). Given the documented concerns regarding the use of chromate and its disposal[7–9], the evolution toward chromate-free corrosion inhibitors is underway. In a tangible sense, there are already numerous chromate-free corrosion inhibition strategies utilised in consumer products today. The adoption of alternative (chromate-free) approaches is progressing as suitable alternatives to chromate are identified albeit few are as (i) cost-effective, (ii) passivating, and (iii) applicable across a wide range of metals and alloys, as chromate.

A review by Gharbi and co-workers into chromate alternatives summarised that singular alternatives to chromate as a 'drop-in' replacement strategy are unlikely[10]. The past three decades have seen much research focus on alternatives for chromates. Some of the most widely explored alternatives that have demonstrated promising performance approaching that of chromate-containing inhibitors include rare-earth-based inhibitors[11] and rare-earth coatings, vanadate-based coatings[12] that are currently utilised in aerospace systems[13,14], lithium-containing coatings[15], organic coatings, nanocomposites, phosphate coatings[16] and metal-rich primers[17]. Undoubtedly, the search for chromate alternates remains a very timely topic and a puzzle that is yet to be solved in the field.

The rapidly growing and large-scale material science knowledge base is typically published as archival 'papers'. In this content, text mining has been one of the most exciting tools in recent years[18–21]. Most literature text remains unstructured or semi-structured data (natural language) which is not capable of being readily interpreted by a computer (whereby a computer is unable to readily interpret context). However, to extract comprehensible and meaningful information from text, supervised natural language processing (NLP) and machine learning methods have been shown to be promising and resulted in the exploration of text mining in the field of material science[22–25]. Supervised NLP requires part of the corpus (i.e., a body of writing) to be in the form of human-annotated data for training, and then tested by unlabelled text. Some supervised NLP algorithms include support vector machines (SVM), bayesian network (BN), maximum entropy (ME), conditional random field (CRF), as well as several other algorithms[26–29]. However, the vast majority—if not essentially all the open literature reports and published data—are unlabelled. Therefore, such text and data may be mined by unsupervised NLP algorithms. Clustering, a well-known unsupervised machine learning algorithm for classifying similar data into groups, has been demonstrated and used to generate machine learning datasets and to identify noisy data, in material science[30,31].

In the present work, the aim was to apply unsupervised NLP in order to explore the suitability of such methods in aiding the

[1]College of Engineering, Computing and Cybernetics, The Australian National University, Acton, Canberra, ACT 2601, Australia. [2]ARC Research Hub for Australian Steel Innovation, Wollongong, Australia. [3]Faculty of Science, Engineering and Built Environment, Deakin University, Waurn Ponds, Geelong, VIC 3216, Australia.
✉email: shujing.zhao@anu.edu.au; nick.birbilis@deakin.edu.au

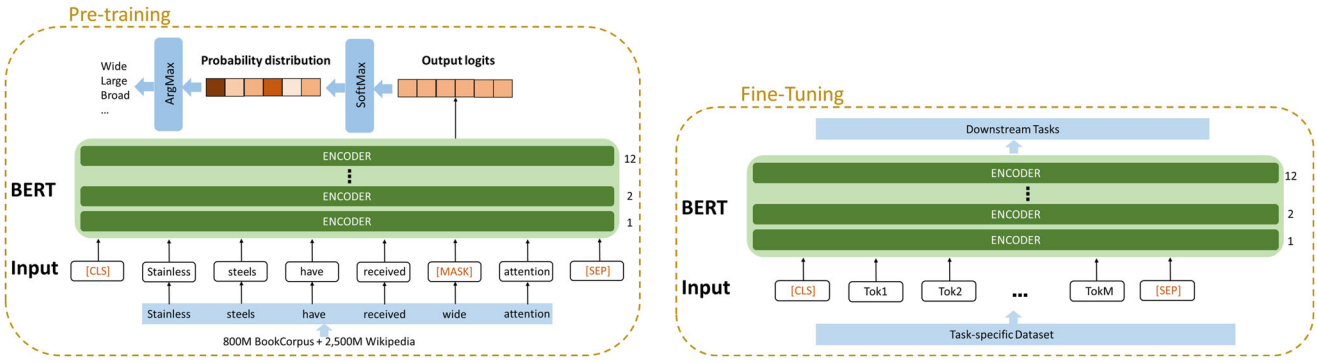

Fig. 1 Schematic representation of Masked language modelling (MLM) with BERT, including pre-training and fine-tuning steps. In the pre-training step, for example, the word 'wide' is randomly masked and special tokens (SEP and CLS) are added as input. The possible predictions for 'wide' in the pre-trained model are 'wide', 'large' and 'broad'. The fine-tuning step takes task-specific datasets as input on the pre-trained model and is designed for various downstream tasks. The number of Transformer encoder layers is L, the hidden size is H, the number of self-attention heads is A. $\text{BERT}_{\text{base}}$ has a model size of $L = 12$, $H = 768$, $A = 12$, total parameter number = 110M.

interpretation of the corrosion-related text; specifically seeking to assess if NLP without the need for a human-in-the-loop could be applied to seeking alternatives for chromates.

Tshitoyan et al.[32] utilised Word2Vec, an unsupervised word embedding method, to extract underlying structure–property relationships in materials and predict new thermoelectric materials. Word2Vec is a vector representation of words, which allows similar words to have a similar representation. Although word embedding is one of the most widely used representations of vocabulary, such an approach can only generate one vector for each word. Therefore, Word2Vec models are context-independent and different contexts of one word are not able to be taken into account. In addition, the Word2Vec model is not capable of learning Out-of-Vocabulary (OOV), since it generates tokens on the 'word' level. However, given the promise that Word2Vec has shown to date, in the field of materials, in this study we apply the method towards the open literature in order to seek possible chromate alternatives. In addition to using Word2Vec, we have also explored the utilisation of a state-of-the-art model, known as bidirectional encoder representations from transformers (BERT).

BERT is a language representation model developed by Google in 2018, enabling pre-training deep bidirectional representations from unlabelled text, by jointly conditioning both the left and right context (outlined further below) in all Transformer encoder layers[33]. In the BERT model, sub-word tokenization is utilised, with a principle that rare words are decomposed into sub-words; whilst frequent words should not be split[33]. This allows the model to process words it has never seen before; meaning BERT is capable of learning Out-of-Vocabulary. The BERT tokenizer is based on WordPiece embedding with 30,000 tokens[33,34], by implementing the following methods: (i) Tokenizing (splitting texts to sub-word tokens), switch tokens to integers, and encoding/decoding; (ii) generating new tokens to the corpus; and (iii) adding and assigning special tokens: the mask to fill (MASK), separator token (SEP) and classification token (CLS).

A commonly used procedure for training models for various tasks in modern NLP systems is to first pre-train a general model on a large amount of unlabelled data, then finetune on downstream NLP tasks including classification, summarisation, etc. Masked language modelling (MLM) is a pre-training method and utilised for how BERT is pre-trained. Taking a sentence, the model randomly masks 15% of the words in the input, then runs the entire masked sentence through the model to predict the masked words[35]. This is different from traditional recurrent neural networks (RNNs) that usually see the words one after the other, or from autoregressive models like generative pre-trained transformer (GPT) that every token only attends to context to its left, which internally masks future tokens[36]. The MLM approach allows the model to learn a bidirectional representation of a sentence. Figure 1 shows how MLM works, the input sentence is first tokenized with special tokens and fed into BERT as a sequence. The pre-training data is 800M words from BooksCorpus and 2500M words from Wikipedia[33].

The representation of every other input word can be weighted by a (attention weight) during learning MASK word. For example, $a = 1$ means that each other word has equal weight in the representation. The tokens are passed to Transformer encoder layers, each layer applies bidirectional self-attention. Inputs then pass through a feed-forward network, then to the next encoder layer. Each output logit is the size of the vocabulary size and is transferred to a probability distribution by applying the softmax function[33]. A softmax function Eq. (1) is a normalisation process that transforms K input values into K values between 0 and 1 which sums to 1.

$$\sigma(\vec{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \tag{1}$$

The output values then can be interpreted as probabilities. Predicted tokens are then calculated by applying argmax to probability distribution. An argmax function Eq. (2) is a function that returns the argument where the function has a maximum value. Given a function, $f{:}X \to Y$, the argmax over subset S of X is defined as

$$\text{argmax}_S f(x) := \{x \in S : f(s) \leq f(x) \text{for all } s \epsilon S\} \tag{2}$$

The argmax is used to locate the token/class with the largest predicted probability. In the case shown in Fig. 1, for example, the predicted results could be 'wide, large, broad, etc.'[37]. The fine-tuning part has a similar architecture to pre-training: For different downstream tasks, feed the model with task-specific inputs, add a task-specific output layer and finetune parameters.

RESULTS

Based on the results from previous work[32] an advantageous outcome from the Word2Vec representation was identified as the ability to represent both 'application words' and 'material formulae' similarly. In the present work, we, therefore, expect the cosine distance of materials that have the same application, to be close; i.e. when the cosine similarity of a material representation and 'chromate' representation is high, such materials are very likely to have similar applications to chromium and therefore candidate alternatives to chromate. Therefore, the closeness of the cosine distance is utilised for the determination of candidate alternatives to chromate when using the Word2Vec model.

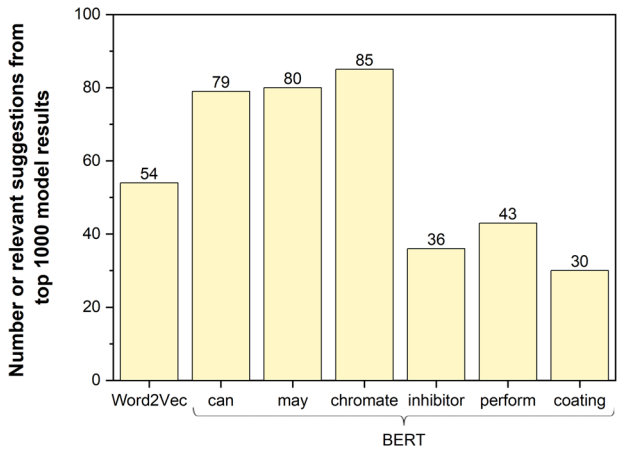

Fig. 2 Number of relevant suggestions that were corrosion protection strategies from the top 1000 results generated by the respective Word2Vec and BERT models. The abbreviation words 'can', 'may', 'chromate', 'inhibitor', 'perform', and 'coating' are six questions that required filling to seek possible suggestions in the BERT model.

In the utilisation of the BERT MLM model, the different mode of operation of the model was exploited in order to seek chromate alternatives by posing open questions—namely six questions that required filling by six different masks. The top predictions arising from the exploration of the six [MASK], were deemed the candidate alternatives for chromate replacements.

For Word2Vec model and each of the six masked BERT models, the top 1000 results generated from the models were extracted, and this list of the top 1000 results was sorted to identify the candidate alternatives that are actually materials (i.e. materials, chemicals, or compounds) and that are relevant to the corrosion domain. Whilst most of the top 1000 results were related to the topic of materials, corrosion, and alloys (to a large extent), we excluded common terms that had no relevance as alternates (i.e. 'steel') and any terms that were not materials (i.e. general conversational words). The number of relevant suggestions for chromate alternatives from the Word2Vec and BERT models, which are highly related to corrosion protection, are shown in Fig. 2.

Of the top 1000 entries generated from each model, the Word2Vec approach detected 54 materials (which is inclusive of materials, compounds, or chemicals) that are relevant to serving as suitable alternatives to chromate. Conversely, the BERT approach identified a number of suitable alternatives that varied from 30 to 85—depending on the question asked. From Fig. 2, it is evident that the first three [mask] containing questions from the BERT model, yielded a relatively higher number of relevant results, which was correlated (by the authors) to how the mask-containing sentence was structured. The BERT model sentences with the lowest yield include 'The best corrosion inhibitor is [mask]' and 'The best conversion coating is [mask]'. These two sentences have the prospect of generating a large number of verbs as outputs, such as 'obtained', or 'needed', or, adjectives as outputs such as 'available', or 'possible'—all of which are reasonable to fill the sentence (but are not relevant materials). If we combine the results from six masked BERT models, the BERT model was capable of predicting 161 individual relevant suggestions for chromate alternatives. A list of each of the ranked chromate alternatives tallied in Fig. 2, is presented in Supplementary Tables 1 and 2.

Of the chromate alternative results predicted by the Word2Vec model, some results were the same as the results from the BERT model, with an overlap of 19%. When interpreting the results obtained, it was observed that the results from Word2Vec have all appeared at least once in the corpus. However, the BERT model was not only able to identify some low-frequency results but also identified results that had never appeared in the corpus. For example, the frequency of 'cvd' (which is chemical vapour deposition) is one and its prediction rank is almost the same as 'nanoparticles' which appears 61 times, as well as, 'formaldehyde', 'acrylate', etc. that do not exist in the dataset. This is an important insight because since the BERT model uses sub-word tokenization during its training, the model can 'mix and match' (between the pre-training and fine-tuning), allowing the model to predict words not seen before in the corpus of corrosion protection relevant training data. The pre-training step, which was carried out using technical documents (from the SciBERT database) and Wikipedia (from the chemBERT database)—is where the BERT model would have seen such words—and is then able to use such words following fine-tuning. This indicates that the ability to pre-train BERT models using a vast array of less-specialist text is meaningful, as the BERT model is able to predict in a human-like manner (including out-of-field).

To illustrate how many alternatives have the potential for the replacement of chromate, we compared the results of the predictions from the Word2Vec and BERT MLM model, with a list of benchmark chromate replacements. The list of potential alternatives was derived from three sources, each of which is a culmination of 'expert' level human analysis—and years of research and literature analysis[10,38,39]. The three studies/reports from which the benchmark list of chromate replacements was derived were not utilised in the model training herein and were reserved as independent validation. The benchmark alternative list was curated into 20 categories, ranging from trivalent chromium, rare-earth-based coatings, vanadate-based coatings, Li-containing coatings, organic systems, and phosphate-based systems to Mg-rich primers—as shown in Table 1.

When reviewing the outputs of the Word2Vec and BERT model, the authors manually identified relevant materials (suitable for consideration as chromate replacements) and allocated them to the benchmark category to which they relate—as also seen in Table 1.

To analyse the efficiency of the NLP models to predict chromate alternatives in an automated manner, we summarised the number of benchmark-related results in each category, for each NLP model and present the results in Fig. 3.

Specifically, a total of 45 results (out of 54 relevant results overall) were considered as relevant benchmark alternatives from the Word2Vec model. The Word2Vec model, therefore, exhibited an 83.3% benchmark-related rate, which was the highest rate—when compared to the six masked BERT models. It is also noted from Fig. 3, that the first three masked sentences (from the BERT model) outperform the latter three masked sentences by a factor of nearly two. The results that do not match benchmark alternatives are either materials/alloys ('magnesium'), substrate materials, or terms that are not materials such as process-related techniques (e.g. 'PVD', 'hard chromium plating'). For example, from the Word2Vec model, such words are epoxy, PVD, hard chromium plating, diamond-like, sol, neodymium, lanthanum, clays, magnesium and Nd.

To investigate these benchmark chromate alternative results in more detail, we focus on benchmark-related results in each category predicted by Word2Vec and BERT model. Figure 4 reveals the count of benchmark-related results in each of the twenty benchmark categories, isolating the performance of the Word2Vec model (in black) and the BERT model (in red). For this analysis, the BERT model is presented as a summation of the six masked models trialled, in order to examine overall performance.

Inspection of Fig. 4 reveals that the Word2Vec model did not identify four categories: trivalent chromium, titanium conversion coatings, zinc-based coatings and calcium-based coatings; while the BERT model covered all 20 categories, with at least one

Table 1. Benchmark alternatives to chromate with given prediction results examples.

#	Benchmark alternatives to chromate[10,38,39]	Examples from the NLP predictions
1	Trivalent chromium	
2	Rare-earth-based inhibitors (aka lanthanide systems)	Cerium, CeN_3O_9
3	Vanadate based inhibitors	BiO_4V ($BiVO_4$), vanadate
4	Li-containing conversion coatings/ primers	Lithium, LDH
5	Organic systems	Polyurethane, amines, BTA
6	Phosphate-based systems	Zinc phosphate, phosphates
7	Mg-rich primers	
8	Zirconium conversion coatings	
9	Titanium-containing conversion coatings	
10	Silicon based systems	Silanes and sol–gel
11	Zinc-based coatings	
12	Molybdate compounds	Molybdate, $MoNa_2O_4$ (Na_2MoO_4)
13	Calcium based systems	
14	Electro-coatings/Electrophoretic systems	Electroplating, Co–P
15	Nanocomposites incl. nanoparticles	Alumina, ceria, titania, graphene
16	TFSAA (thin film sulfuric acid anodisation)	Anodising
17	BSAA (boric sulfuric acid anodising)	Anodising, sulfuric
18	Filled electroless nickel	
19	WC/C (tungsten carbide carbon) coating	WC–Co, WC, $(WCr)_2C$-Ni
20	Nitride coatings	TiN, Cr_xN, TiN/CrN

The list of benchmark categories was derived from three research sources[10,38,39]. The prediction results in both Word2Vec and BERT models were manually identified, and part of the relevant materials was allocated here.

prediction. One of the categories, 'silicon-based systems', showed the highest number of predictions by the Word2Vec model. The other category of 'organic systems', revealed the same number of predictions by the BERT model. To further probe the four categories which were only identified by the BERT model, we report the prediction materials in each category and their frequency of occurrence in the original dataset, listed in Table 2. The frequency with which most of these materials were mentioned was relatively high, ranging from 25 to nearly 450 instances. We found that these words never appeared explicitly in the same sentence with 'chromate', but they connected to 'chromate' through other ways, such as 'hydroxide' occurs in the same paragraph with 'Cr', 'TCP' and 'chromium'.

Overall, whilst the Word2Vec model and BERT model revealed the highest benchmark-related rate, that rate is only one metric of performance—and is directly linked to the number of predicted results. One of the more holistic assessments of model performance, when inspecting Fig. 4, suggests that the BERT model outperformed the Word2Vec model for detecting all of the 20 benchmark chromate alternatives—including in the variety of approaches therein. Whilst not necessarily probed further in the present work, the ability of the BERT model to predict chromate replacement results, was also described. Conforming to the prediction of benchmark results—whilst meaningful in the initial exploration and validity of NLP approaches—is a strong

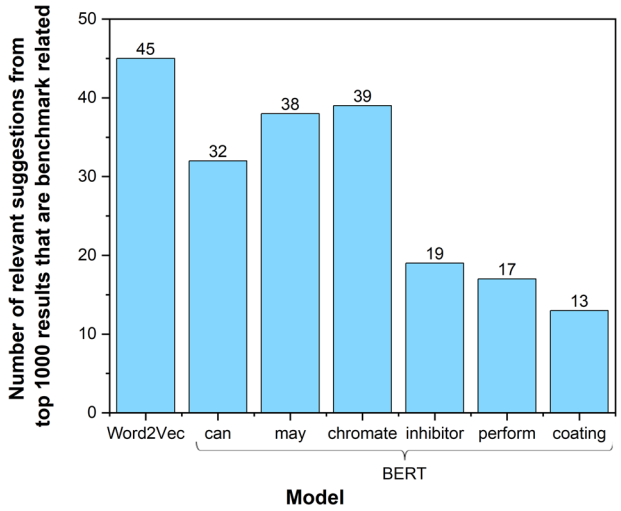

Fig. 3 Number of benchmark alternatives materials in corrosion-related results. The abbreviation words 'can', 'may', 'chromate', 'inhibitor', 'perform' and 'coating' are six questions that required filling to seek possible suggestions in the BERT model.

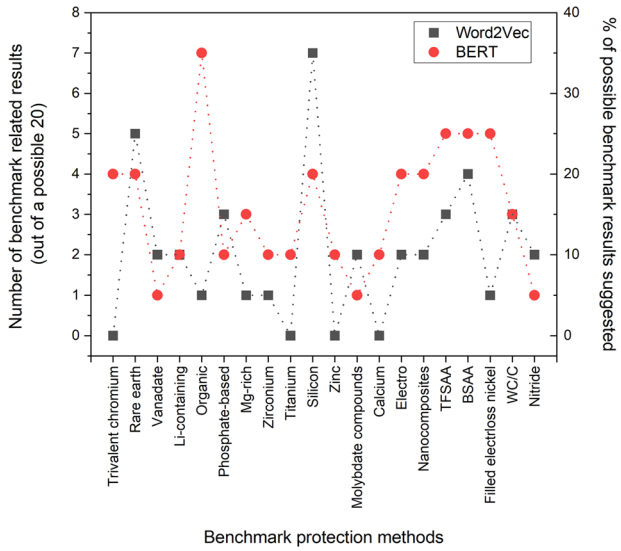

Fig. 4 Number of benchmark-related predictions by Word2Vec and BERT model in 20 benchmark alternative categories. In each benchmark category, the benchmark-related results of Word2Vec model (in black) and BERT model (in red) were counted. The BERT model herein is a summation of the six filling mask models.

confirmation that expert human-level interpretation is capable (in an automated process). From an aspirational perspective, the NLP approaches should extend beyond human-level benchmarking and identify results and correlations that not readily having been interpreted by humans.

DISCUSSION

When comparing the predictions of six masked BERT models, it was noted that the input-masked sentence structure significantly affects the performance of outputs. Since our goal herein was to search for possible chromate replacement, output noise data such as common words, verbs or adverbs should be avoided to the maximum extent. As a consequence, and based on initial work herein, it is posited that when the input masked word is a noun, this possibly correlates with better prediction performance

Table 2. Four benchmark alternatives categories that only identified by BERT model.

#	Category	Benchmark alternatives	Frequency
1	Trivalent chromium	Fluoride, hydroxide, HF, ceramic	25/52/24/291
2	Titanium conversion coatings	Titanium, Ti	300/253
3	Zinc-based coatings	Zinc, Zn	298/443
4	Calcium based systems	Calcium, Ca	87/83

Benchmark alternatives are prediction results identified by the six filling mask BERT models but not in the Word2Vec model. Frequency is their number of occurrences in the original dataset.

(when comparing with expert human-level benchmarks). As a way to analyse how the top 1000 chromate alternative predictions have identified corrosion-protection-relevant results, we identified the corrosion-relevant results from the Word2Vec and BERT models. It is acknowledged that the comparison between the two is not apples-and-apples; as the Word2Vec model runs as one/singular analysis; whereas the BERT model may be re-run numerous times (herein we used six masked models overall, and could have used more). This feature of the models tends to favour the BERT model, for 'user experience' and its ability to further probe for results. In regards to the latter point, here we revealed that the BERT model yielded more relevant results by almost three times, compared to the Word2Vec model. This finding suggests the BERT model, instead of focusing on the word 'chromate', learns from context, builds instance-specific embedding for sentences, and therefore has a wider interpretation of corrosion protection-relevant materials. However, when examining if the relevant results from the models tested are correlating with the benchmark-related materials (Table 2), the Word2Vec model outperformed BERT, which is attributed to its word-sense for 'chromate' and any known chromate replacement is expected to be 'highly related' to 'chromate' and thus, have a similar word embedding. Whilst there is an ability to therefore capture what is already known as expert-human-level chromate alternatives from the literature using NLP (as demonstrated herein, and indeed by the Word2Vec model), this means that NLP can alleviate the need for a human to read and interpret hundreds (if into thousands) of papers. The task of becoming an expert can be replicated. However, the next step in suggesting alternatives for chromate, and nuance in context-related interpretations, could be seen in the performance of the BERT model. The BERT model was able to predict certain benchmark alternatives that were not closely surrounded by 'chromate' in the dataset. Additionally, BERT was capable of predicting low-frequency and even zero-frequency alternatives—known as out-of-bag words. That is because BERT can generate vector representation in different sentences and there are infinite embeddings for each word type (and some of the out-of-bag words were indeed candidate alternatives). The study herein was an initial demonstration of NLP in the context of a corrosion challenge, and also, one of the first utilisations of the BERT model in a practical/applied engineering problem. Based on the work herein, there is significant scope for broadening the corpus for training, to include patents, websites, and other works—beyond the Scopus API, in the search for chromate alternatives.

In this study, natural language processing (NLP) was utilised to automate the search of scientific literature for chromate replacements; specifically in the context of corrosion protection. It was revealed that the application of NLP was capable of serving in the role of searching for chromate replacements, without the need for a human to read any of the associated scientific literature. Herein, two NLP approaches were utilised, namely, the Word2vec approach (previously explored in the field of materials by others) and the BERT approach, recently developed by Google. The latter approach was explored on the basis of its potential in handling out-of-vocabulary words, and its ability to operate by finding alternative words for a [mask] (i.e. the ability to 'answer questions asked' of the BERT model). The finding from the study herein can be summarised as

- When comparing the NLP predictions from the work herein (which did not have a human in the loop) with three (3) benchmark studies/reviews from corrosion experts that have proposed a list of chromate replacements, it was determined that:

  The Word2vec model predicted the most accurate chromate alternative results, by simply calculating the cosine distance.
  The BERT model predicted the most extensive related results in the field, inclusive of even low-frequency terms.
  Both the BERT and Word2vec models could capture essentially all of the expert human-determined chromate replacement technologies—albeit with no domain experience.
- NLP was able to readily capture scientific knowledge for a niche application, revealing the approaches employed herein—not developed for the application of chromate replacement—can serve as general approaches for broad applications.
- This study presented a descriptive model for summarising chromate replacements from the literature without human annotation, by using NLP. This is a first-attempt report focused on insight into the past and aims to identify materials that experts can identify by extracting existing corrosion knowledge.
- Future work may explore the use of broader inputs, beyond those of the Scopus application programming interfaces, including webpages, and other collections. Future work may explore the use of materials properties, corrosion and protection mechanisms. Specifically, broadening the inputs and including mechanistic facets will possibly permit more chromate replacement predictions that are not in the benchmark alternative list.

METHODS

Data collection and pre-processing

A total of 5990 entries were collected by accessing and extracting 84 million records from Scopus application programming interfaces (APIs) (https://dev.elsevier.com/). A set of wild card query terms was introduced to limit acquisition primarily related to the relevant topic. Only articles with 'chrom*'and 'replace*' or 'substitute' in their titles, abstracts or keywords were collected. Furthermore, abstracts were filtered by applying query terms 'alumin*', 'zinc', 'magnesium', 'alloy', 'steel' or 'iron' (to ensure that they were relevant to substrates of interest). Abstracts that were in non-English languages were removed from the corpus to allow the use of a singular language setting as English. A number of articles with copyright limitations or missing passages were also removed, as were articles with content types not corresponding to peer-reviewed publications—leaving 1812 works forming the training dataset for the Word2Vec and BERT architecture.

Preprocessing of body text involved removing XML format and XML quotes tags, leading words such as 'Abstract' were also eliminated. In the Word2vec model, we followed the general preprocessing steps as per the unsupervised word embedding study from 2019[32]. Element and element names, numbers and units were converted to tokens, such as #element, #nUm, #unit, respectively. Material formulas were normalised in an alphabetical way, such that any chemical formula was simplified regardless of

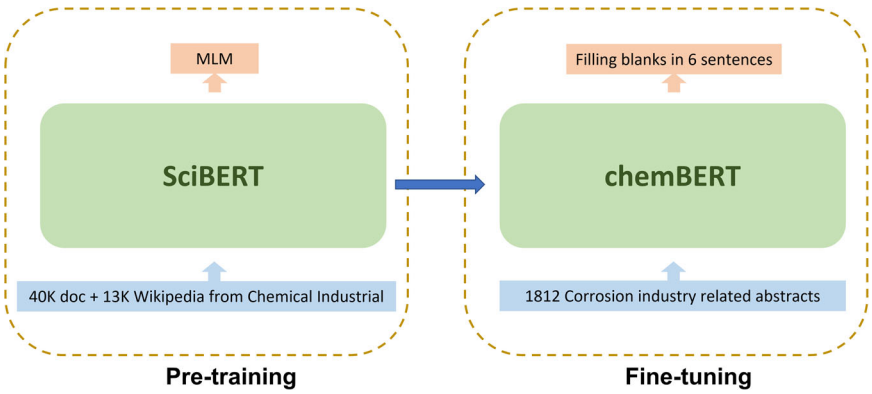

Fig. 5 The pre-training and fine-tuning of BERT model in this study. The chemBERT model was pre-trained on the SciBERT model (designed for scientific texts) with technical documents and Wikipedia from the Chemical Industry. The chemBERT model was then finetuned using corrosion-related text and was used to fill the mask in six sentences.

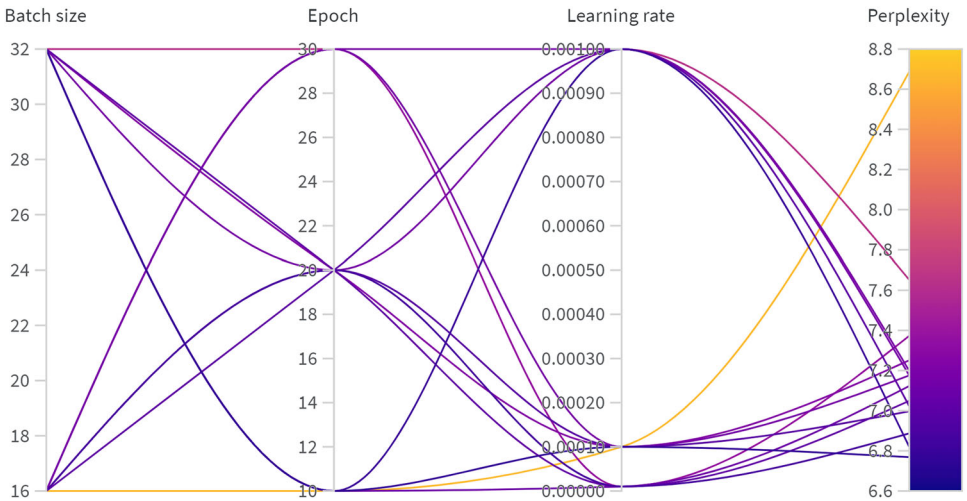

Fig. 6 Hyperparameter tuning in BERT model to find the best hyperparameters for training. Each hyperparameter set (batch size, epoch and learning rate) was trained on a development dataset (10% of the whole dataset). The lower perplexity shows a darker purple colour, indicates better performance of the language model with certain pair of hyperparameters.

the order of elements. The processed dataset includes one study in each line, and it was tokenized to a combination of individual words through ChemDataExtractor[40]. Chemical formulas were recognised by applying pymatgen[41], regular expression and rule-based techniques, jointly. The body text was transformed to lowercase if the token was not a chemical formula or an abbreviation. Abbreviations were identified by instances when not only the first letter was uppercase. In the pre-trained BERT-based model, subword tokenization was used allowing the model to process words it has never seen before. The tokenizer includes 250,000+ tokens from the chemical domain. Rare words were tokenized into meaningful subwords while frequent words were only split into word tokens. BERT takes the whole input as a single sequence. Special tokens [CLS] and [SEP] were used to understand the input sequence. Besides token embeddings, BERT includes more information for each token with positional embeddings and segment embeddings.

Training

In the Word2vec training, the same gensim model was utilised, following the hyperparameter tuning process performed on 14,042 material science analogy pairs, as shown in Tshitoyan's work[32]. Hyperparameters are substantial parameters that control the learning process and are evaluated prior to the training. Common hyperparameters include learning rate, optimisation method, loss function, number of hidden layers, batch size, and epochs. Batch size is the length of data samples for training before the gradient descent updates. A development dataset for the hyperparameter tuning process was created herein, in which 10% of data were extracted randomly from the original dataset. The optimisation process is a grid search and searches through a specified set of parameters in the hyperparameter space. Models were trained with each pair hyperparameters and evaluated by the evaluation metric: analogy score, as discussed in the subsequent Evaluation section. A set of optimal hyperparameters were gained with the highest analogy score: a learning rate of 0.001, a size of embedding of 300, and a batch size of 128 and 30 epochs. The training is then performed by applying the set of optimal hyperparameters.

To focus the study in the Chemistry Domain, we used the pretrained chemical-bert-uncased model in Hugging Face[40], and finetuned the training based on the mask language modelling implementation in Huggin Face with some modifications[42]. This pretraining and finetuning process is shown in Fig. 5. The chem-bert-uncased model is pretrained from SciBERT (https://huggingface.co/allenai/scibert_scivocab_uncased) with over 40,000 technical documents from Chemical Industrial and over 13,000 Wikipedia Chemistry articles. The software 'Wandb' (for

Table 3. The six masked sentences designed to seek potential alternatives for chromate in the BERT model.

#	Masked sentence	Abbreviation
1	hexavalent chromium can be replaced by [MASK]	can
2	hexavalent chromium may be replaced by [MASK]	may
3	chromate can be replaced by [MASK]	chromate
4	the best corrosion inhibitor is [MASK]	inhibitor
5	[MASK] performed better than chromate	perform
6	the best conversion coating is [MASK]	coating

Chromate replacements were identified by filling mask in the finetuned BERT model. [MASK] is a 'blank' for the model to predict, this approach is known as Fill Mask. Six abbreviations represent six Fill Mask tasks.

tracking weights and biases) was introduced to track and visualise the hyperparameter tuning process[43]. Similarly, the best hyperparameters for finetuning were selected by training a model on a small parcel of data (the development set) over each pair hyperparameter and calculating corresponding perplexity, as discussed in Evaluation. The hyperparameters pairs are epoch = (10,20,30), batch size = (16,32), learning rate = ($1e^{-5}$, $1e^{-4}$, $1e^{-3}$). As shown in Fig. 6, the best perplexity corresponds to the optimal hyperparameter pair (epoch = 10, batch size = 32, learning rate = $1e^{-4}$). We then fine-tuned the model with this optimal hyperparameter pair on the processed abstract dataset.

The fine-tuning process of BERT is notionally considered 'straightforward' for various downstream tasks (e.g. classification, sequence labelling and question answering). By adding one or more additional layers after the final pre-training layer, it is typical to freeze the early BERT layers and only train the later layers. Such downstream tasks are usually performed on task-specific labelled texts, and therefore most of the fine-tuning processes in BERT are supervised. However, the fine-tuning in the present study is still unsupervised; instead using masked language modelling with non-labelled corrosion-related text. Pretraining on domain-specific data in NLP tends to yield higher performance[44–46]. Therefore, we apply pre-training on chemical domain data (chemical-bert-uncased model[40]) and fine-tuning with corrosion domain information (via the Scopus API) to strengthen the understanding of the language model on chromate replacement. The application of masked language modelling is predicting which word is filled in the sentence, which is defined as 'Fill Mask'.

Evaluation

NLP tasks generally can be validated with measures of accuracy including *f*-score, root mean squared error (RMSE), etc. However, the evaluation of unsupervised learning can be challenging due to the unlabelled output. This is primarily because common evaluation methods require comparing an output value against a known value. For the Word2vec model, the evaluation metric is the analogy score, defined as the rate of correctly matched analogies from two chemical and element name pairs. Usually, the evaluation metrics for MLM are cross entropy and perplexity. Perplexity (given by Eq. (3)) is a commonly used value to evaluate language models in NLP:

$$PP(W) = 2^{H(W)} = 2^{-\frac{1}{N}\log_2 P(w_1, w_2, \ldots, w_N)} \quad (3)$$

where H is the cross-entropy, P is the language model, w is a sequence of words, and N is the length of the words. A lower perplexity commonly indicates a better language model with more predictable results.

Herein, the means by which the Word2Vec and BERT models identify chromate replacement materials were designed. In the Word2Vec study, the cosine distance to vector 'chromate' was used, to represent the probability of a material being a chromate replacement. That is, the chromate replacements are among the materials most similar to chromate, which were determined by the projection of normalised word embeddings. While in the BERT experiment, we attempted to identify chromate replacements by 'filling blanks' in a sentence, this approach is known as Fill Mask. For example, top predictions of potential chromate replacements were sought by filling a [mask], whereby an example is: 'Chromate can be replaced by [mask]'. Our model randomly masks 15% of the input, runs the whole masked sentence, and outputs the prediction of the masked words. Both the predicted results were categorised and compared with a benchmark alternative list which is summarised from known alternative corrosion preventative technologies, as discussed in the Results. The six masked sentences fed into the BERT model to seek potential alternatives to chromate are listed in Table 3. The 'can', 'may', 'chromate' and 'perform' sentences were designed to explore the model providing direct answers for possible chromate replacement materials, while the 'perform' sentence had distinctly different structure and embedded comparison semantics. The 'inhibitor' and 'coating' sentences, on the other hand, were designed according to the main application of chromate—whereby the top corrosion inhibitor and conversion coating materials were also seen as potential alternatives.

Herein, the alternatives to chromate predicted were refined to a list of 20 categories of benchmark alternatives, as described in the "Results" section.

DATA AVAILABILITY

The abstracts used in this study are available via Elsevier's Scopus API's (https://dev.elsevier.com/). The list of DOIs and all other data generated and analysed in the current study, are available from the corresponding authors on reasonable request.

CODE AVAILABILITY

The Word2vec word embeddings are adapted from https://github.com/materialsintelligence/mat2vec. The fine-tuning process of BERT model is adapted from https://github.com/huggingface/transformers/tree/main/examples/pytorch/language-modelling. The data collection and pre-processing, hyperparameter tuning, model evaluation and all other code in the current study, are available from the corresponding authors on reasonable request.

Received: 15 August 2022; Accepted: 21 December 2022;
Published online: 06 January 2023

REFERENCES

1. Koch, G. et al. International measures of prevention, application, and economics of corrosion technologies study. *NACE Int.* **216**, 2–3 (2016).
2. Hou, B. et al. The cost of corrosion in China. *npj Mater. Degrad.* **1**, 1–10 (2017).
3. Resona Ltd. *Impact of Corrosion in Australasia Report* (The Australian Corrosion Association, 2021).
4. IARC. *Some Inorganic and Organometallic Compounds. Monographs on the Evaluation of the Carcinogenic Risk of Chemicals to Man* Vol. 2 (International Agency of Research in Cancer, Lyon,1973).
5. Frankel, G. S. & McCreery, R. L. Inhibition of Al alloy corrosion by chromates. *Electrochem. Soc. Interface* **10**, 34 (2001).
6. IARC. Chromium, nickel and welding. IARC monographs on the evaluation of carcinogenic risks to humans. *Int. Agency Res. Cancer* **49**, 49–256 (1990).
7. Smith, E. & Ghiassi, K. Chromate removal by an iron sorbent: mechanism and modeling. *Water Environ. Res.* **78**, 84–93 (2006).
8. Saha, R., Nandi, R. & Saha, B. Sources and toxicity of hexavalent chromium. *J. Coord. Chem.* **64**, 1782–1806 (2011).
9. Pellerin, C. & Booker, S. M. Reflections on hexavalent chromium: health hazards of an industrial heavyweight. *Environ. Health Perspect.* **108**, A402–A407 (2000).
10. Gharbi, O., Thomas, S., Smith, C. & Birbilis, N. Chromate replacement: what does the future hold? *npj Mater. Degrad.* **2**, 1–8 (2018).
11. Hinton, B. Corrosion inhibition with rare earth metal salts. *J. Alloy. Compd.* **180**, 15–25 (1992).
12. Guan, H. & Buchheit, R. Corrosion protection of aluminum alloy 2024-T3 by vanadate conversion coatings. *Corrosion* **60**, 284–296 (2004).

13. Kiyota, S., Valdez, B., Stoytcheva, M., Zlatev, R. & Schorr, M. Electrochemical study of corrosion behavior of rare earth based chemical conversion coating on aerospace aluminum alloy. *ECS Trans.* **19**, 115 (2009).
14. Hamdy, A. S., Doench, I. & Möhwald, H. Vanadia-based coatings of self-repairing functionality for advanced magnesium Elektron ZE41 Mg–Zn–rare earth alloy. *Surf. Coat. Technol.* **206**, 3686–3692 (2012).
15. Visser, P. et al. The corrosion protection of AA2024-T3 aluminium alloy by leaching of lithium-containing salts from organic coatings. *Faraday Discuss.* **180**, 511–526 (2015).
16. Weng, D., Jokiel, P., Uebleis, A. & Boehni, H. Corrosion and protection characteristics of zinc and manganese phosphate coatings. *Surf. Coat. Technol.* **88**, 147–156 (1997).
17. King, A. & Scully, J. Sacrificial anode-based galvanic and barrier corrosion protection of 2024-T351 by a Mg-rich primer and development of test methods for remaining life assessment. *Corrosion* **67**, 055004-055001–055004-055022 (2011).
18. Tan, A.-H. Text mining: the state of art and the challenges. In *Workshop on Knowledge Discovery from Advanced Databases (KDAD'99)* 71–76 (1999).
19. Hotho, A., Nürnberger, A. & Paaß, G. A brief survey of text mining. In. *LDV Forum* **20**, 19–62 (2005).
20. Hassani, H., Beneki, C., Unger, S., Mazinani, M. T. & Yeganegi, M. R. Text mining in big data analytics. *Big Data Cogn. Comput* **4**, 1 (2020).
21. Berry, M. W. & Kogan, J. *Text Mining: Applications and Theory* (John Wiley & Sons, 2010).
22. Friedman, C., Kra, P., Yu, H., Krauthammer, M. & Rzhetsky, A. GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. In *ISMB (Supplement of Bioinformatics)* 74–82 (2001).
23. Müller, H.-M., Kenny, E. E., Sternberg, P. W. & Ashburner, M. Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biol.* **2**, e309 (2004).
24. Kim, E. et al. Materials synthesis insights from scientific literature via text extraction and machine learning. *Chem. Mater.* **29**, 9436–9444 (2017).
25. Leaman, R., Wei, C.-H. & Lu, Z. tmChem: a high performance approach for chemical named entity recognition and normalization. *J. Cheminform.* **7**, 1–10 (2015).
26. Rameshbhai, C. J. & Paulose, J. Opinion mining on newspaper headlines using SVM and NLP. *Int. J. Electr. Comput. Eng. Syst.* **9**, 2152–2163 (2019).
27. Sohn, S. et al. Detection of clinically important colorectal surgical site infection using Bayesian network. *J. Surg. Res.* **209**, 168–173 (2017).
28. Berger, A., Della Pietra, S. A., & Della Pietra, V. J. A maximum entropy approach to natural language processing. *Comput. Linguist.* **22**, 39–71 (1996).
29. Sutton, C. & McCallum, A. An introduction to conditional random fields. *Found. Trends Mach. Learn.* **4**, 267–373 (2012).
30. Parker, A. J. & Barnard, A. S. Selecting appropriate clustering methods for materials science applications of machine learning. *Adv. Theory Simul.* **2**, 1900145 (2019).
31. Li, H. et al. Clustering discretization methods for generation of material performance databases in machine learning and design optimization. *Comput. Mech.* **64**, 281–305 (2019).
32. Tshitoyan, V. et al. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* **571**, 95–98 (2019).
33. Jacob Devlin, M.-W. C., Kenton L, & Toutanova, K. Bert: pre-training of deep bidirectional transformers for language understanding. In *Proc. NAACL* 4171–4186 (2019).
34. Schuster, M. & Nakajima, K. Japanese and Korean voice search. In *2012 IEEE International Conference on Acoustics, Speech And Signal Processing (ICASSP)* 5149–5152 (2012)
35. Horev, R. BERT Explained: state of the art language model for NLP. *Towards Data Sci.* **10**, https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270 (2018)
36. Radford, A., Narasimhan, K., Salimans, T. & Sutskever, I. Improving language understanding by generative pre-training. *OpenAI.* https://www.cs.ubc.ca/~amuham01/LING530/papers/radford2018improving.pdf (2018)
37. Mani, S. P. et al. Effect of multilayer CrN/CrAlN coating on the corrosion and contact resistance behavior of 316L SS bipolar plate for high temperature proton exchange membrane fuel cell. *J. Mater. Sci. Technol.* **97**, 134–146 (2022).
38. Wiley, B. *REACH Compliant Hexavalent Chrome Replacement for Corrosion Protection (HITEA).* Technology Strategy Board Project 101281 (Technology Strategy Board, 2014).
39. Pollard, D. *Chromate-Free Coatings Systems for Aerospace and Defence Applications* https://pra-world.com/2019/08/21/chromate-free-coatings-systems-for-aerospace-and-defence-applications/ (2019).
40. Swain, M. C. & Cole, J. M. ChemDataExtractor: a toolkit for automated extraction of chemical information from the scientific literature. *J. Chem. Inf. Model.* **56**, 1894–1904 (2016).
41. Ong, S. P. et al. Python Materials Genomics (pymatgen): a robust, open-source python library for materials analysis. *Comput. Mater. Sci.* **68**, 314–319 (2013).
42. Wolf, T. et al. Transformers: state-of-the-art natural language processing. In *Proc. 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* 38–45 (2020).
43. Biewald, L. *Experiment Tracking with Weights and Biases* https://www.wandb.com/ (2020).
44. Gu, Y. et al. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthc.* **3**, 1–23 (2021).
45. Sung, C. et al. Pre-training BERT on domain resources for short answer grading. In *Proc. 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* 6071–6075 (2019).
46. Møller, A. G., Van Der Goot, R. & Plank, B. NLP North at WNUT-2020 task 2: pre-training versus ensembling for detection of informative COVID-19 English Tweets. In *Proc. Sixth Workshop on Noisy User-generated Text (W-NUT 2020)* 331–336 (2020).

ACKNOWLEDGEMENTS

We gratefully acknowledge the funding from the Australian Research Council (ARC) through the Industrial Transformation Research Hubs Scheme under Project Number: IH200100005.

AUTHOR CONTRIBUTIONS

All authors contributed to the design of the study, as well as the manuscript, S.Z. collected the abstracts and performed data processing, trained and evaluated the Word2Vec and BERT model, and drafted the paper. N.B. set up query terms for accessing the corpus, set up benchmark alternatives table, modified plots, as well as reviewed, edited and modified the manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41529-022-00319-0.

Correspondence and requests for materials should be addressed to Shujing Zhao or Nick Birbilis.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.