www.nature.com/scientificreports

scientific reports

OPEN

Tracking 35 years of progress in metallic materials for extreme environments via text mining

Xin Wang[1], Anshu Raj[2], Yanqing Su[3], Shuozhi Xu[2] & Kun Lu[1]✉

As global energy demands rise, the advancement of new energy technologies increasingly relies on the development of metals that can endure extreme pressures, temperatures, and fluxes of energetic particles and photons, as well as aggressive chemical reactions. One way to assist in the design and manufacturing of metals for the future is by learning from their past. Here we track the progress of metallic materials for extreme environments in the past 35 years using the text mining method, which allows us to discover patterns from a large scale of literature in the field. Specifically, we leverage transfer learning and dynamic word embeddings. Approximately one million relevant abstracts ranging from 1989 to 2023 were collected from the Web of Science. The literature was then mapped to a 200-dimensional vector space, generating time-series word embeddings across six time periods. Subsequent orthogonal Procrustes analysis was employed to align and compare vectors across these periods, overcoming challenges posed by training randomness and the non-uniqueness of singular value decomposition. This enabled the comparison of the semantic evolution of terms related to metals under extreme conditions. The model's performance was evaluated using inputs categorized into materials, properties, and applications, demonstrating its ability to identify relevant metallic materials to the three input categories. The study also revealed the temporal changes in keyword associations, indicating shifts in research focus or industrial interest towards high-performance alloys for applications in aerospace and biomedical engineering, among others. This showcases the model's capability to track the progress in metallic materials for extreme environments over time.

Keywords Metallic materials, Extreme environments, Dynamic word-embedding, Semantic evolution, Transfer learning

Achieving the urgent societal goals of reduced emissions and increasing energy efficiency is driving the development of new metallic materials capable of exceptional performance in extreme environments[1]. For example, metals that can withstand high pressure and corrosion are vital for underwater applications. These properties are crucial for both deep-sea exploration and maintaining the integrity of subsea pipelines[2]. In the arena of space exploration, alloys that resist radiation and drastic temperature fluctuations are indispensable for the durability and functionality of spacecraft and satellites[3]. Similarly, for defense applications, the development of more effective protective gear and armaments depends on metals that can survive the intense heat and pressure from ballistic impacts[4]. Moreover, in the realm of renewable energy, metallic materials that endure the harsh conditions faced by wind turbines and solar panels are key to improving the longevity and efficiency of these sustainable energy technologies[5]. These examples underscore the importance of research into advanced metals designed for extreme environments, which is fundamental to technological progress and environmental stewardship[6].

Future design and manufacturing processes of metals can be aided by drawing insights from their historical development. For example, the recent development of the CrTaVW and CrHfTaVW refractory multi-principal element alloys (MPEAs) for irradiation environments[7,8] was in part motivated by the much earlier finding that the element W possesses high melting temperature, low plasma erosion rate, reduced tritium retention, and ability to refine grains. On the one hand, one main source of useful information for metallic materials is the scientific literature. On the other hand, the sheer size and growth of the literature present challenges for researchers to gain an up-to-date, comprehensive understanding of the historical trends of any specific field. Therefore, it is critical to efficiently sift through vast volumes of literature to extract, categorize, and synthesize

[1]School of Library and Information Studies, University of Oklahoma, Norman, OK 73019, USA. [2]School of Aerospace and Mechanical Engineering, University of Oklahoma, Norman, OK 73019, USA. [3]Department of Mechanical and Aerospace Engineering, Utah State University, Logan, UT 84322, USA. ✉email: kunlu@ou.edu

relevant data by using techniques such as text mining. In this way, we can uncover valuable patterns, trends, and correlations that inform and inspire the next generation of metal design and manufacturing innovations.

The fields of text mining and natural language processing have witnessed rapid development in the last decade[9]. One of the most notable innovations is the development of large scale self-supervised language modeling, such as word embedding, which shows an impressive understanding of the underlying semantics of human languages. The evolution of word embedding models started with static embeddings, exemplified by Word2Vec[10]. These models assign each word a single, fixed vector representation derived from its overall context of occurrence. Subsequently, context-aware embeddings were introduced, including GloVe[11], BERT[12], and GPT[13]. Unlike Word2Vec, these models generate context-dependent word representations that adapt based on the specific context in which a word appears. Only more recently, the word embedding technique has made into the materials science domain. In 2019, Tshitoyan et al.[14] trained a word embedding model on 1.5 million abstracts related to inorganic materials, enabling the recommendation of functional materials years before their discovery. Zhang and He[15] presented an unsupervised machine learning (ML) approach that utilizes Word2Vec to predict solar cell materials from textual data in scientific literature, demonstrating its efficacy with first-principles validation. Huang et al.[16] introduced BatteryBERT, a pre-trained language model tailored for enhancing battery research databases by efficiently processing and extracting information from scientific texts. More recently, Pei et al.[17] extended the word embedding model to metallic materials using 6.4 million abstracts to capture the similarity of chemical elements among 2.6 million candidate high entropy alloys (HEAs).

While language is acknowledged to evolve over time[18], these static word embedding models are not suitable to reveal the evolving relationships among material concepts. Training separate models on time-sliced data is possible, but will lead to incomparable models because the word vectors from different time slices may not align to the same coordinate axes in the high dimensional vector space[19]. To address this, we aligned the models trained from time-sliced data to a benchmark word embedding model, making them comparable across different periods. This alignment is crucial, as the static word embedding model is theoretically unfit for the dynamic nature of this study. Orthogonal Procrustes analysis was conducted to align these embeddings with the benchmark word embeddings, thereby enabling comparisons across different time periods. To the best of our knowledge, no study has used dynamic word embedding (DWE) models to unveil the trend of materials studies over time. DWEs extend the word embedding models to capture the evolving word semantics over time, thereby being suitable for tracking the evolution of a field.

Therefore, inspired by these previous successes of text mining in other material problems[14,17], we create a DWE model to study the trends of metallic materials for extreme environments (Fig. 1). Firstly, abstracts and titles from nearly one million articles spanning from 1966 to 2023 were downloaded from the Web of Science, with 1966 chosen as the starting point because it is the beginning year in the database's records on "metallic materials for extreme environments." Building upon the model by Pei et al.[17], these abstracts and titles were used for transfer learning to obtain benchmark word embeddings. This method has been proven effective in grasping domain nuances[20]. Since the literature before 1989 only had titles but no abstracts in the Web of Science database, the years from 1989 to 2023 were divided into six time periods, and new DWE were trained using word2vec.

Results

Trends in broad topics

To analyze the evolution of research topics over time, we utilized the Latent Dirichlet Allocation (LDA) topic modeling[21], which is a widely used probabilistic algorithm for uncovering latent topics in text data. The choice

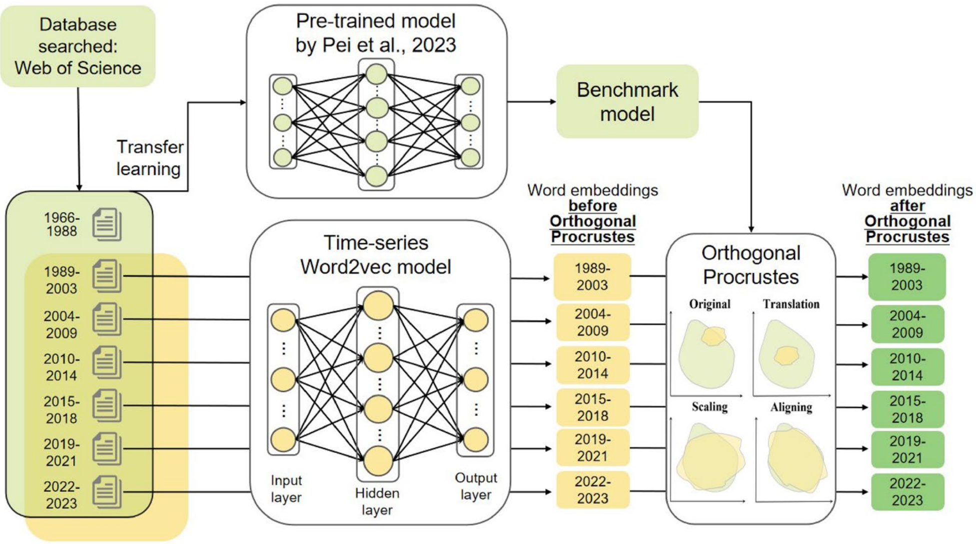


Fig. 1. Schematic for our DWE model.

of LDA was motivated by its effectiveness in handling large collections of unstructured text data and its ability to identify underlying patterns that evolve over time. Unlike other topic modeling techniques, LDA's hierarchical Bayesian framework ensures that topics are generated with minimal bias[21], making it suitable for exploratory analyses across diverse research domains.

The parameters for the LDA model were selected to balance interpretability and model stability. The number of topics was set to 7 based on domain expert knowledge. `no_above` was set to 0.5 to filter out common terms that appear in more than 50% of the documents, ensuring the topics remained specific and meaningful. Similarly, `passes` was set to 15 and `iterations` to 400 to allow the model sufficient time to converge, thus producing stable topic distributions. These values were determined empirically, as increasing them further yielded diminishing returns in topic quality while significantly increasing computational time.

Supplementary Table S1 presents the raw output from the LDA model. To enhance the coherence and interpretability of the generated topics, we reviewed the results and consolidated the seven topics into three themes for each time period. This step addressed occasional inconsistencies in the topics, such as unclear groupings of terms or significant overlap between topics.

Table 1 illustrating the topic trends from 1989 to 2023 provides an overview of the evolving research interests over the years. From 1989 to 2003, the focus was predominantly on chemically active environments, high/low temperatures, and thin film materials. This period marked the beginning of research into environments where chemical reactions played a crucial role, as well as the study of materials under extreme temperature conditions and the properties of thin films. Between 2004 and 2009, the research landscape began to shift slightly. The research interest is still focused on chemically active environments but also on high pressures. Thin film materials continued to be a significant topic which is the ongoing research in this field. From 2010 to 2014, environmental science emerged as a key area. This reflects a growing awareness of environmental issues. Chemically active environments remained relevant, and biological science appeared as a new focus, indicating an interest in the biological applications of these studies. The period from 2015 to 2018 showed a strong emphasis on environmental science. Studies on chemically active environments persisted and high/low temperatures reappeared as a significant topic. In the period of 2019–2021, battery research emerged prominently, reflecting the global push towards sustainable energy solutions. Environmental science continued to be a vital area of research, along with ongoing studies on chemically active environments. The most recent period, 2022 to 2023, further highlighted the growing focus on battery technology and innovation. Environmental science continued to be of importance, and research into chemically active environments continued unabated. This analysis of topic trends underscores the dynamic nature of research interests, with emerging fields such as environmental science and battery technology gaining prominence while maintaining a consistent focus on chemically active environments.

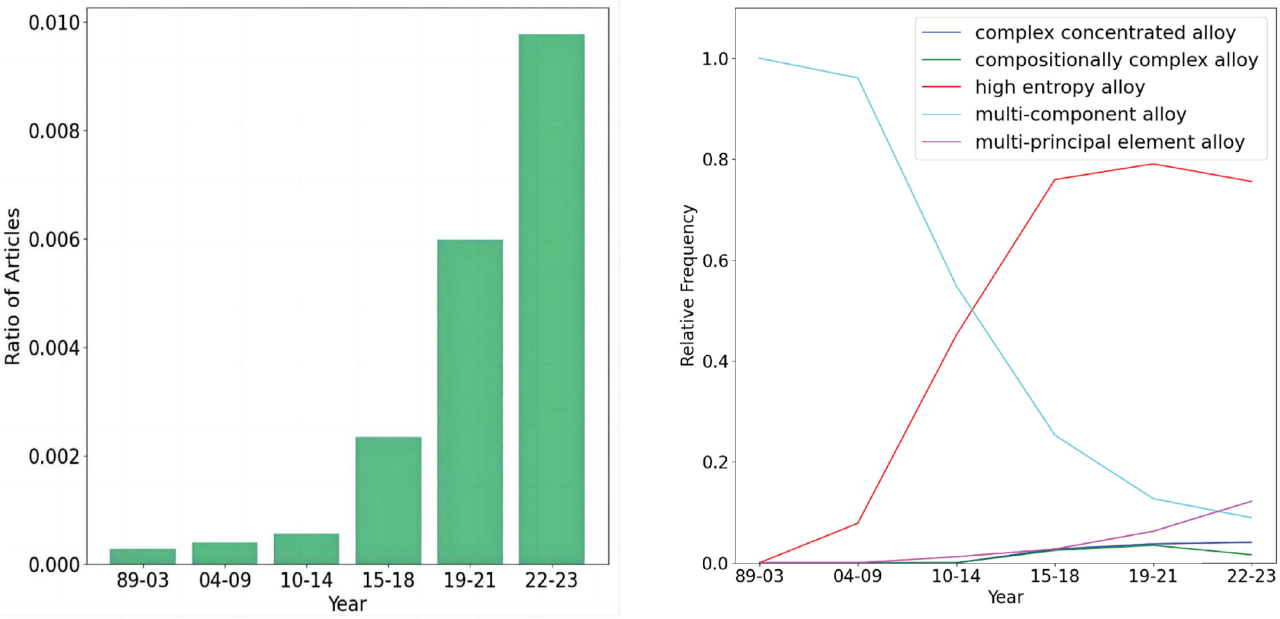
Trends in research on a specific type of materials for extreme environments

Figure 2 presents a trend analysis based on the frequency of keyword occurrences in the scientific literature on a specific type of material for extreme environments.

Figure 2a shows the ratio of articles containing at least one of the five specified keywords ("high entropy alloy", "multi-principal element alloy", "multi-component alloy", "complex concentrated alloy", "compositionally complex alloy") to the total number of articles collected in each time period. The *x*-axis represents time periods, while the *y*-axis denotes the ratio of articles. A noticeable trend is observed in which the ratio of articles containing these keywords significantly increases over time, starting from nearly negligible in the 1989–2003

Period	Topics
1989–2003	Chemically active environments
	High/low temperatures
	Thin film
2004–2009	Chemically active environments
	High pressures
	Thin film
2010–2014	Environmental science
	Chemically active environments
	Biological science
2015–2018	Environmental science
	Chemically active environments
	High/low temperatures
2019–2021	Batteries
	Environmental science
	Chemically active environments
2022–2023	Batteries
	Environmental science
	Chemically active environments

Table 1. Topic trends by year range in the context of "metallic materials for extreme environments".



(a) Ratio of articles over time

(b) Relative frequency of keywords over time

Fig. 2. Evolution of research interest. (**a**) The ratio of articles is equal to numbers of articles containing at least one of the five specified keywords to the total number of articles in each time period. (**b**) The y-axis is calculated by dividing the number of articles containing a specific keyword by the number of articles with at least one of the five keywords in the same time period.

period to about 10% in the latest 2022–2023 period. This indicates a growing interest and research focus on these specific types of alloys in the context of extreme environments.

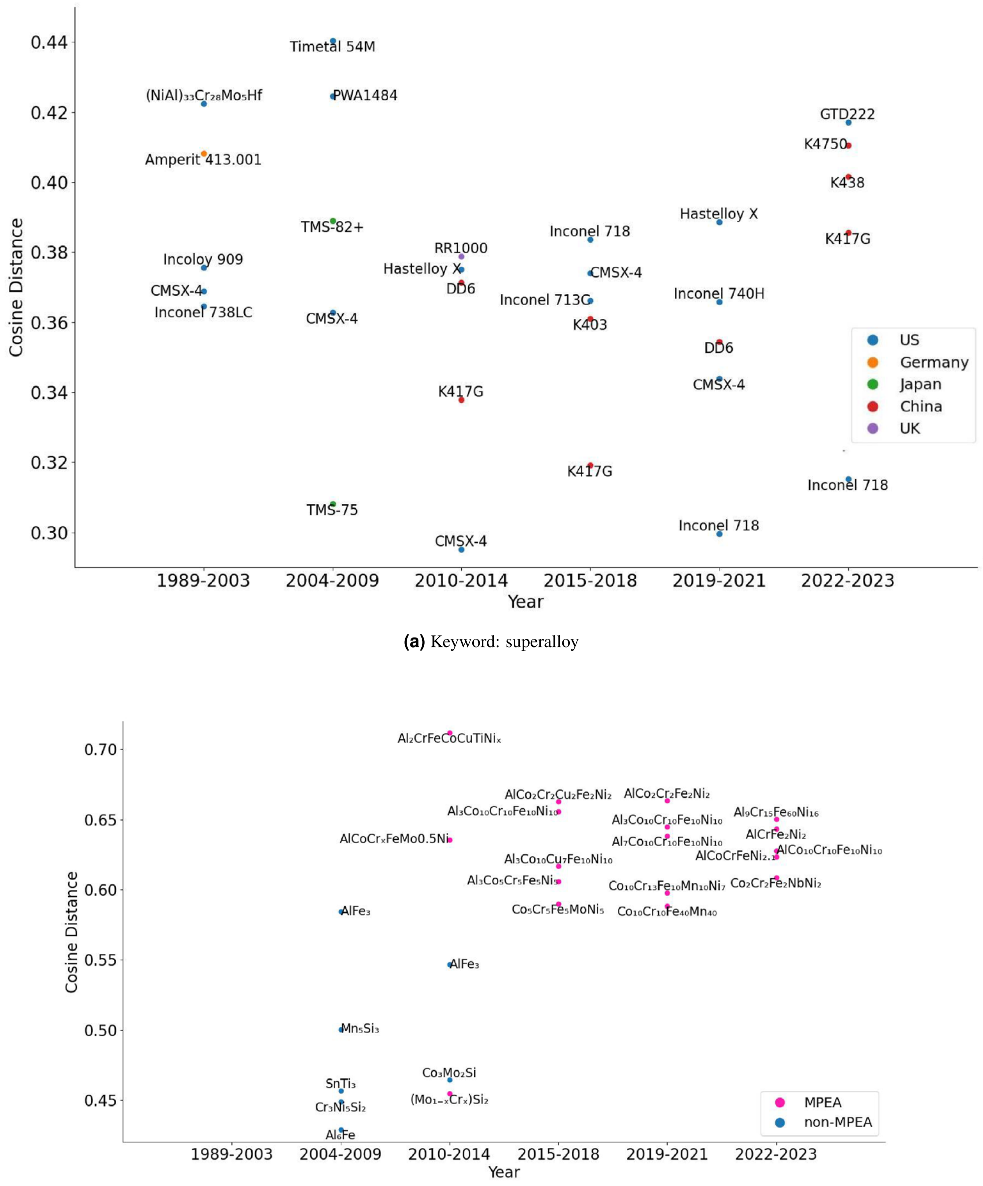
Figure 2b is a line chart that illustrates the relative frequency of each keyword over the same six time periods. The *x*-axis is again the time periods, and the *y*-axis measures the relative frequency of each keyword. This is calculated by dividing the number of articles that contain a specific keyword by the number of articles that have at least one of the five keywords in the same time period.

In the provided graph, the first emergence of "high entropy alloys" was observed in 2004-2009. This depicted trend is consistent with the fact that the concept of HEA was first proposed in 2004[22,23]. Then research into these alloys significantly accelerated in the 2010s, marking an era of intensive exploration and application of HEAs[24–26]. Correspondingly, "multi-principal element alloy", "complex concentrated alloy", and "compositionally complex alloy" saw a moderate rise and maintained a relatively steady presence throughout the periods. The trend aligns with the fact that researchers gradually realized that many MPEAs (e.g., ternaires and quaternaries) do not have a sufficiently high configurational entropy of mixing and thus cannot be called "high entropy alloys"[27]. Lastly, the term "multi-component alloy" exhibits a steep decline after dominating the earlier periods. Historically, the term was used to refer to all alloys that contain more than two chemical elements without specifying their entropy or elemental concentration, so it covers dilute alloys as well. One explanation for the decrease in frequency of the term is that, with the emergence of HEAs, researchers began to differentiate between HEAs, medium-entropy alloys, dilute alloys, and other categories. Hence, ambiguous terms such as "multi-component alloy" became less popular.

Evolution of keyword-metal relationship

Figures 3, 4, and 5 show how the association between specific keywords and metals evolves over time. In each figure, the *x*-axis represents different time intervals, while the *y*-axis indicates cosine distance. The closer the distance, the higher the cosine similarity between the material word vectors and the input keywords, indicating a stronger association. Given the complexity of chemical material nomenclature, where multiple names may refer to the same substance, manual annotation has been applied to the model's output to ensure more accurate representation and interpretation.

In Fig. 3a, all tagged superalloys are Ni-based, and no Co-based alloys are included. This shows the dominance of Ni-based superalloys. When labeling alloys by their country of origin, we found that from 2010, Chinese superalloys like "DD6" and "K417" began to emerge. The cosine similarity between these Chinese materials and "superalloy" indicates a high level of contextual similarity in the literature. This means that these material names are more likely to appear in similar contexts as "superalloy" in the abstracts we collected during this period. When labeling alloys by their country of origin, we found that from 2010, Chinese superalloys like "DD6" and "K417G" emerged. Since 70% of superalloys are used in aerospace engineering[28], our finding is in line with China's growing aerospace engineering sector, as evidenced by the rapidly increasing export market trade volume in high-tech aerospace since 2007[29]. Figure 3b reveals a marked prevalence of the elements Al, Co,



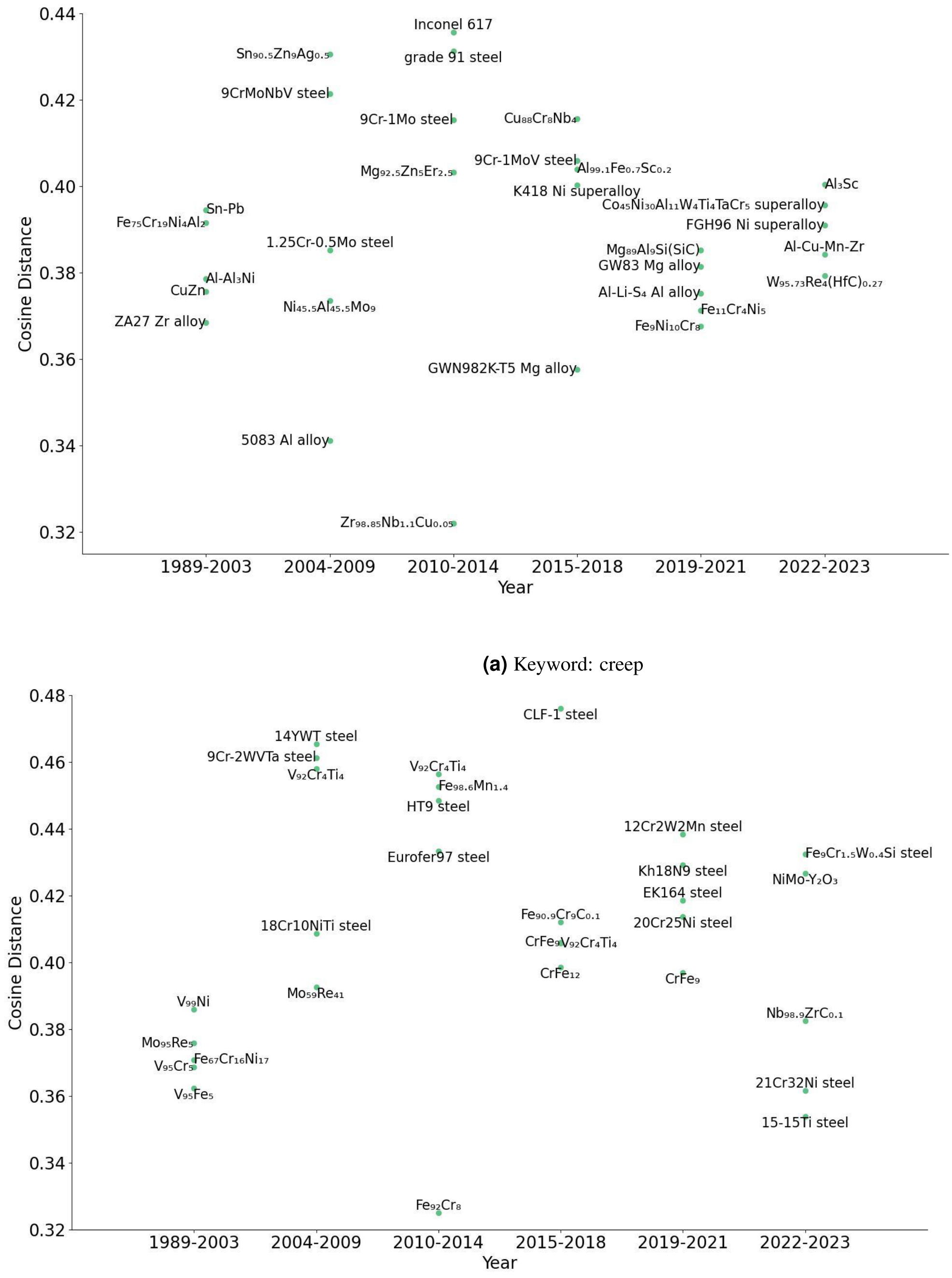
(a) Keyword: superalloy

(b) Keywords: high_entropy_alloy, multi-principal_element_alloy, multi-component_alloy, complex_concentrated_alloy, compositionally_complex_alloy

Fig. 3. Input a type of material to obtain specific materials.

Cr, Cu, Fe, and Ni in HEAs/MPEAs. This observed trend aligns with the prior findings in all HEAs/MPEAs, e.g., Figure 2 of Ref.[30] and Figure 7 of Ref.[27]. Specifically, although the Figure 2a indicates that "multi-component" appeared in abstracts and titles between 1989 and 2003, it was not recognized as a distinct phrase by the model due to its low threshold score. Such consistency validates the predictive accuracy of our model and suggests the likelihood of these materials maintaining their prominence in the near future. Additionally, when we input multiple keywords, the model identifies words that are closely related to the average vector of all the entered keywords. This process explains the presence of some non-MPEAs, especially in 2004–2009.

To demonstrate our model's capability to identify potentially significant alloys ahead of their time, we utilized the same keywords as shown in Figure 3b to search for materials using data spanning from 2004 to 2009. This time period is notable for two reasons: (i) the concept of HEA was first proposed in 2004[22,23], and (ii) the first refractory HEAs were developed in 2010[31]. A domain expert screened the list of metals returned by our model and identified Ti-Nb-Zr (ranked 222) and Ti-13Nb-13Zr (ranked 381) as refractory MPEA. Tracing these alloys, we found a 2004 paper[32] developed three alloys — Ti-13Nb-13Zr, Ti-20Nb-13Zr, and Ti-20Nb-20Zr. Thermodynamic calculations using the TCHEA7 database in Thermo-Calc[33] predicted the liquidus temperatures of these alloys as 1668°C, 1698°C, and 1681°C, respectively. All exceed the melting point of Ti, the primary constituent element, whose liquidus temperature is 1667°C. Notably, the 2004 study[32] did not use the term HEA or MPEA. However, our model successfully associated these alloys with these concepts, suggesting



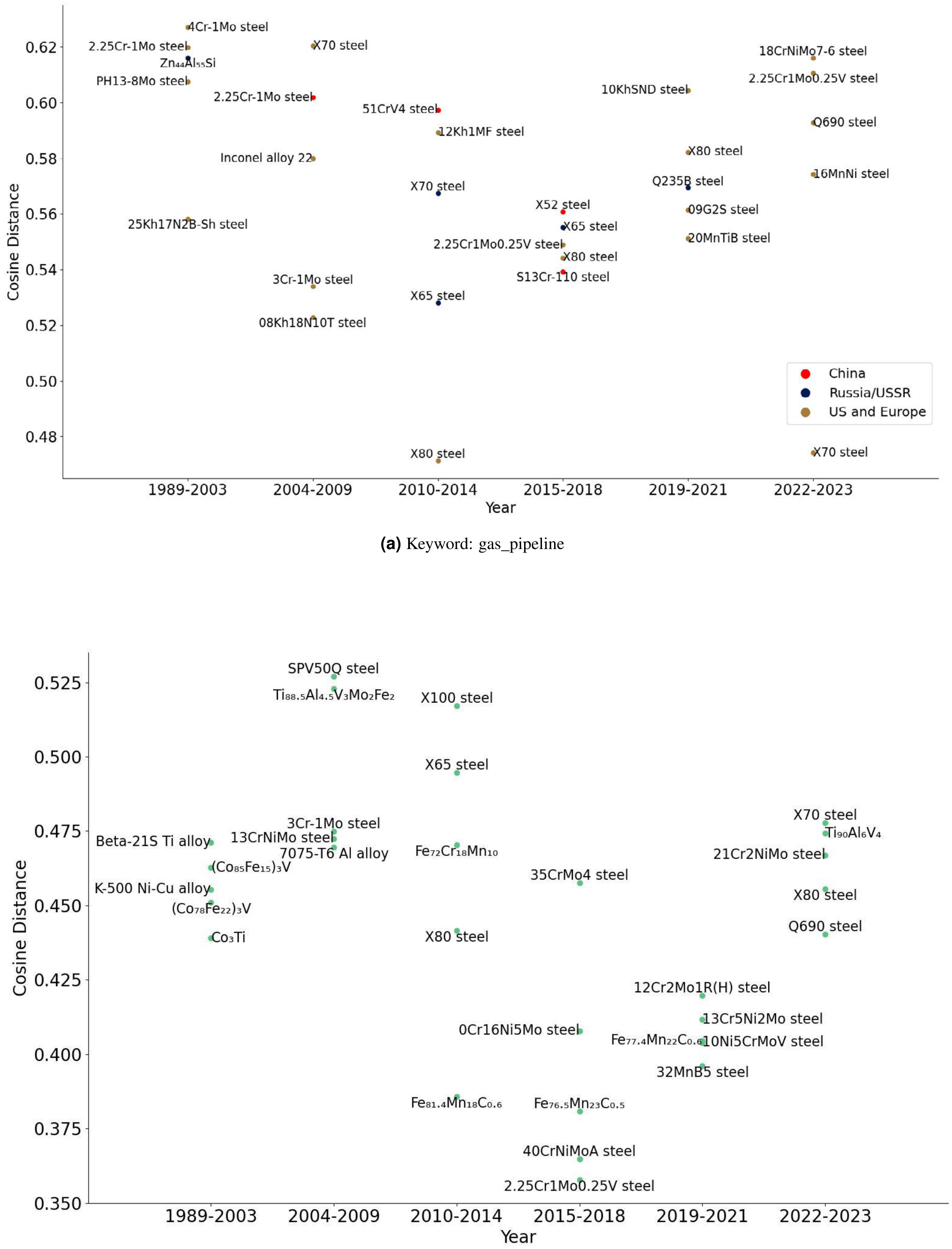
(a) Keyword: creep

(b) Keywords: neutron_irradiation, proton_irradiation

Fig. 4. Input properties to obtain specific materials.

that researchers were exploring refractory MPEAs prior to the formalization of HEA terminology in the same year. Historically, this indicates that the emergence of "refractory HEA" or "refractory MPEA" was inevitable.

Figure 4a shows that earlier materials associated with creep contained the element Pb, while no alloys after 2004 contained it anymore, reflecting the fact that Pb failed out of favor in structural alloys due to its toxicity[34]. Another trend over time is that while steels remain popular, there's a visible shift towards Ni and Co based superalloys. This shift suggests an evolving preference for superalloys as the go-to materials for creep resistance, indicating advancements in materials science and heightened performance requirements in relevant applications[35]. Figure 4b illustrates a shift in materials associated with irradiation: initially, V alloys were common, but since 2019, the focus has shifted almost to various types of steels. This change likely reflects



(a) Keyword: gas_pipeline

(b) Keyword: hydrogen_embrittlement

Fig. 5. Input applications to obtain specific materials.

research identifying disadvantages in V alloys, such as swelling and transmutation under irradiation. The current trend points towards the sustained preference for steel-based materials in radiation-resistant applications such as nuclear reactors.

Figure 5a shows an initial diversity of materials, including Ni and Zn alloys, which then exclusively shifted to steels in more recent years, possibly due to their improved resistance to hydrogen-related material challenges. Since 2019, Chinese steels have surfaced in the dataset, coinciding with China's ascent as a major steel producer, accounting for over half of the global output from 2018[36]. This trend may point to the growing prevalence of Chinese steels in gas pipeline applications moving forward. Figure 5b suggests a similarity with Fig. 5a, showing a predominance of steels. This trend likely reflects the importance of hydrogen embrittlement resistance in materials used for gas pipelines, a logical emphasis given their use in hydrogen gas transport.

In most cases, as time progresses, the association between some metals and keywords exhibits periodic fluctuations. This reflects the alternating changes in technological advancement, industrial demand, and research trends. For these alloys, their exceptional ability to attain excellent mechanical properties even in extreme environments might be the reason for their frequent appearances across multiple time periods. For instance, when the keyword "superalloy" is searched, the CMSX-4 alloy first appears in the top 5 list in the 2004–2009 period; then it is considered the dominant alloy from 2010 to 2014; it follows that CMSX-4 appears in the top 5 with reduced relevance in later years; eventually, it disappears from the list of the top 5 alloys. This phenomenon indicates that research interest in specific alloys may increase due to the discovery of new applications and then decrease as more advanced materials emerge.

Implications for future research

Studying the trends of metallic materials in extreme environments allows us to review significant achievements, identify emerging trends, and recognize patterns and relationships between different concepts. This analysis provides valuable insights into the development and optimization of materials. It highlights technological advancements, innovative approaches, and key factors influencing material performance. This facilitates the discovery of new materials and the improvement of existing ones for applications in challenging environments. For example, increased interest in MPEAs/HEAs, shown in Fig. 2a, might stimulate deeper research into the physical properties of such alloys; attention to specific properties may promote the improvement of related alloys to meet the needs of specific industrial applications.

Our findings have significant implications for future developments. Regarding trends in material types, Figs. 4b and 5b indicate an increasing preference for steels in environments exposed to irradiation or gaseous hydrogen. Additionally, Fig. 4a illustrates a growing utilization of superalloys for creep resistance applications, while the use of steels for this purpose is declining. Concerning the trends in the materials' countries of origin, Figs. 3a and 5a respectively suggest that, if current patterns persist, superalloys and steels manufactured in China will become increasingly predominant. Understanding these trends can inform future materials selection and guide strategic decision-making in engineering applications.

Conclusion

This study advances the understanding of metallic materials in extreme environments using transfer learning and orthogonal Procrustes analysis. It builds on Pei et al.'s model[17], enhancing it with a corpus from Web of Science, featuring nearly 1 million titles and abstracts. The corpus, covering the literature from 1989 to 2023, was divided into six periods and processed through a skip-gram word2vec model to generate time-series word embeddings. Despite achieving accurate models for each period, aligning these embeddings posed challenges due to the inherent randomness in training and the non-uniqueness of SVD solutions. Orthogonal Procrustes analysis was employed to align vectors from different periods, facilitating temporal comparison of word semantics. The study uses cosine similarity within an 8-word context window to analyze word semantics, focusing on the evolution of terms related to metallic materials under extreme conditions. Three input types (materials, properties, and applications) were tested to demonstrate the model's effectiveness in identifying specific chemical materials, with a noise-management filtering mechanism refining the results.

Future research could build upon the findings of this study by leveraging the identified trends in material evolution to predict and develop next-generation metallic materials tailored for extreme environments. The demonstrated growth in interest for materials like HEAs/MPEAs suggests an opportunity to explore their application in emerging industries such as renewable energy and advanced aerospace systems. Future studies could focus on integrating predictive modeling with DWE to anticipate material behaviors and properties before experimental validation. This research could also explore cross-disciplinary applications, integrating insights from environmental science and biomedicine to design alloys with multifunctional capabilities for diverse extreme conditions.

Limitations

The chemical terminology recognition and processing mechanisms in this study have certain limitations. During our preprocessing phase, we observed instances where multiple expressions refer to the same substance. For example, Inconel 718 (US), GH4169 (China), and UNS N007718 (international) refer to the same Ni-based superalloy. In this work, different names were manually merged into one. In this sense, further improvements in preprocessing are needed, particularly in the recognition and processing of chemical material terminology, to ensure accurate and consistent automatic handling of such cases. A knowledge base such as an ontology or a knowledge graph covering the field of metallic materials can organize these synonyms in concepts and help identify their relationships. One such example is Wikidata. Future study will incorporate Wikidata to merge these synonyms. The other limitation of this study is that we did not use full text. Currently, major publishers such as Elsevier and Springer Nature have permitted full-text access for text-mining research on subscribed content. Provided that our institutions subscribe to an article, the inclusion of full text will reveal more details about the trends of the field.

Methods

Data collection and preprocessing

To obtain text in a digital format suitable for text mining, we used the Web of Science database as the source for data collection. A boolean search formula was designed to search for literature related to "metallic materials in extreme environments." Keywords and search expressions were carefully selected to ensure the search results cover the field's representative research achievements as comprehensively as possible. The design of the search formula and the search results are shown in Table 2. We manually downloaded all results by clicking on "FAST 5000" on the search results page from the Web of Science database. This button allows us to download relevant information from up to 5,000 articles at once. Once the data was downloaded, we primarily used the title and

Retrieval Parameters	Retrieval Details
Database Searched	Web of Science
Search Formula	ALL=((“metal*” or “alloy*” or “metallic material*”) and (“extreme condition*” or “extreme environment*” or “severe condition*” or “harsh condition*” or “harsh environment*” or “harsh service condition*” or “harsh service environment*” or “extreme operating condition*” or “high temperature*” or “high-temperature” or “cryogenic temperature*” or “cryogenic-temperature” or “low temperature*” or “low-temperature” or “high pressure*” or “high-pressure” or “creep*” or “shock*” or “impact*” or “high-strain-rate” or “ballistic” or “ablation” or “vaporization” or “neutron*” or “proton*” or “helium” or “photon” or “photons” or “radiation*” or “irradiation*” or “nuclear reactor*” or “fission reactor*” or “fusion reactor*” or “cladding material*” or “corrosion*” or “corrosive” or “molten-salt” or “molten salt*” or “oxidation” or “oxidative” or “hydrogen embrittlement” or “chemical reduction” or “chemically reactive”))
Document Type	Articles, Proceeding Papers, Book Chapters, etc.
Time Span	1966–2023
Number of Results	≈ 1 million
Search Content	Year of Publication, Titles, and Abstracts

Table 2. Strategies to acquire research data.

abstract fields for text mining. We employed the method of Tshitoyan et al.[14] to preprocess the corpus. The main preprocessing steps include the following:

(i) Convert text into a list of tokens, making adjustments specific to materials science, such as separating units from numbers and element symbols from their valence states.
(ii) Apply various optional transformations to a series of tokens or a string: excluding punctuation, writing numbers in their standard form, standardizing material formulas, removing accents, and combining tokens into phrases. This method prepares text data for further analysis or ML tasks in materials science.
(iii) Transform chemical formulas into a standardized form, with elements ordered alphabetically and stoichiometric values adjusted to the smallest common integer denominators.

In the model construction, phrase expressions (e.g., “high_entropy_alloy”) were detected using the threshold score. This score of a phrase (i.e., two consecutive words $word_a$ and $word_b$) can be represented by the following formula:

$$\text{score}(word_a, word_b) = \frac{\text{count}(word_a, word_b) - \text{min\_count}}{\text{count}(word_a) \times \text{count}(word_b)} \times N \tag{1}$$

$\text{count}(word_a, word_b)$ is the number of times the word pair $word_a$ $word_b$ co-occurs in the corpus. min_count is a parameter within the model used to ignore words and word pairs that appear less than this value. $\text{count}(word_a)$ and $\text{count}(word_b)$ are the number of times $word_a$ and $word_b$, respectively, appear in the entire corpus. N is the total number of word occurrences in the corpus.

However, not all phrases that are of interest were successfully detected by the threshold method due to their low co-occurrence frequency in our corpus. To address this issue, we compile a list of crucial phrases related to material properties, manufacturing processes, and applications that should be emphasized. These phrases, along with those from Wikipedia’s “List of named alloys,” are searched in our corpus. When found, they are connected with underscores to be treated as phrases. Tokenizing specific phrases within the corpus better maintains their original semantic integrity, aiding the model in more effectively learning the relationships between words, especially in scenarios with limited training data.

Benchmark and DWE

As we train multiple word embedding models over different time periods, we need to align the resulting models in the same vector space to make them comparable. A benchmark word embedding model was built for this purpose. The benchmark model is preferably to be representative and comprehensive of the domain. Because our corpus only represents a portion of the metallic materials-related literature, we use transfer learning to construct a more comprehensive benchmark model. Transfer learning enhances model adaptability by leveraging knowledge from previously trained related domains, simplifying the recalibration process and addressing challenges related to data scarcity and obsolescence[37]. This approach is crucial for updating embedding models over time. The word-embedding model from Ref.[17] is chosen as the starting point for this benchmark model. Then, we update this model by using titles and abstracts of all years after preprocessing as the corpus for fine-tuning. This enables the model to better align with the topics related to metallic materials in extreme environments.

Since the Web of Science database did not present abstracts (but only titles) for the literature prior to 1989, we constructed DWE using the corpus from 1989 onwards. To avoid issues such as inconsistent model quality, excessive noise, and model bias, we divided the corpus into six time periods based on publication years: 1989–2003, 2004–2009, 2010–2014, 2015–2018, 2019–2021, and 2022–2023. The number of articles in these periods were 157451, 128026, 146382, 171475, 186900, and 141124 respectively.

All word embeddings are obtained by building skip-gram word2vec models and are represented as vectors. First, a vocabulary is built on the text after preprocessing and phrase detection. Then, each word in the corpus is treated as the target. To maintain consistency with the model designed by Pei and others, we set the context window size to 8, meaning that up to 8 words before and after the target word are considered as context. The model predicts words within this context based on the target word, transforming the target word into a 200-dimensional vector. The model then goes through the entire training set 30 times for comprehensive learning and adjustment. Through iterative learning, it continuously adjusts its internal vector representations to minimize prediction errors. The training objective of this Skip-gram model is to find word representations. More formally, given a sequence of training words $w_1, w_2, w_3, \ldots, w_T$, the objective of the Skip-gram model is to maximize the average log probability[38]:

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0}\log p(w_{t+j} \mid w_t) \tag{2}$$

where T is the length of the training word sequence, c is the size of the context window, w_t is the center word, and w_{t+j} is the context word adjacent to the center word. However, the basic Skip-gram formula using the softmax function to define $p(w_{t+j} \mid w_t)$ incurs high computational costs, as the cost of computing the gradient is proportional to the vocabulary size W. To address this issue, Hierarchical Softmax is introduced as an efficient approximation method. Hierarchical Softmax simplifies computations and leverages word frequency information to significantly accelerate training. It transforms the problem of comparing a large vocabulary into a decision sequence with a depth of $\log_2(V)$, where V is the size of the vocabulary. This approach effectively reduces the computational load, thereby speeding up the training process.

Orthogonal procrustes analysis

Relatively accurate word vectors were obtained through iterative training. However, randomness in the training process and the non-uniqueness of SVD cause embeddings trained at different times to misalign. This means that even the same word might have different representations in different embedding spaces. To enable comparison of the vector of the same word across different time periods, orthogonal Procrustes analysis is used. This analysis finds an optimal orthogonal transformation (involving translation, scaling, and aligning) to ensure that vectors are aligned on the same coordinate axis, thereby making it possible to compare word vectors across different time periods. X and Y represent the sets of word vectors from the benchmark model and models in different time periods corresponding to the common vocabulary. Through Procrustes analysis, an orthogonal matrix O is found to minimize the difference between X and the transformed Y, given by:

$$\min_{\mathbf{O}\in\mathcal{O}_\mathbf{d}} \|\mathrm{X}-\mathrm{YO}\|_\mathrm{F}^2 \tag{3}$$

where $\|\cdot\|_\mathrm{F}$ denotes the Frobenius norm, and $\mathcal{O}_d$ is a set of orthogonal matrices. The standard approach to solving this optimization problem involves computing the SVD of $\mathrm{Y}^\mathrm{T}\mathrm{X}$, i.e.,

$$\mathrm{Y}^\mathrm{T}\mathrm{X} = \mathrm{U}\Sigma\mathrm{V}^\mathrm{T} \tag{4}$$

The optimal O is then given by:

$$\mathrm{O} = \mathrm{UV}^\mathrm{T} \tag{5}$$

Finally, all word vectors from models in different time periods are transformed using the orthogonal matrix O, and these transformed vectors are then normalized:

$$\mathrm{Y}_{\mathrm{aligned}} = \frac{\mathrm{YO}}{||\mathrm{YO}||} \tag{6}$$

Similar metals filtering mechanism

The model learns the meanings of words by studying their context in the literature, operating under the assumption that words with similar contexts have similar meanings. It transforms words into vector forms, which can represent the semantic information of words in high-dimensional space. For a given set of *n* keywords, their corresponding word vectors are w_1, w_2,...w_n. The following method is used to traverse all vectors in the model's vocabulary space and find the words closest to the average vector of these keywords:

$$\text{cosine\_similarity}\left(\frac{1}{n}\sum_{i=1}^{n}\mathrm{w}_i, \mathrm{v}\right) = \frac{\left(\frac{1}{n}\sum_{i=1}^{n}\mathrm{w}_i\right)\cdot\mathrm{v}}{\left\|\frac{1}{n}\sum_{i=1}^{n}\mathrm{w}_i\right\|\,\|\mathrm{v}\|} \tag{7}$$

Test	Input Type	Input Keywords
1	Materials	superalloy
2	Materials	high_entropy_alloy, multi-principal_element_alloy, multi-component_alloy, complex_concentrated_alloy, compositionally_complex_alloy
3	Properties	creep
4	Properties	neutron_irradiation, proton_irradiation
5	Applications	hydrogen_embrittlement
6	Applications	gas_pipeline

Table 3. Input types and keywords for tests.

$$\text{cosine\_distance}\left(\frac{1}{n}\sum_{i=1}^{n}\mathrm{w}_i, \mathrm{v}\right) = 1 - \text{cosine\_similarity}\left(\frac{1}{n}\sum_{i=1}^{n}\mathrm{w}_i, \mathrm{v}\right) \quad (8)$$

where v represents any other word vector. Cosine distance is reported in this paper for term associations.

To demonstrate the models' performance, the output results are evaluated for accuracy from three dimensions in Table 3: (i) Materials — The first and second tests focus on finding specific material types based on input keywords such as "high entropy alloy", "multi-principal element alloy", "complex concentrated alloy", "compositionally complex alloy", and "superalloy"; (ii) Properties — The third and fourth tests identify and recommend specific materials based on input property keywords, such as creep, neutron irradiation, and proton irradiation; (iii) Applications — The last two tests determine and recommend appropriate materials based on specific application scenarios, such as hydrogen embrittlement and gas pipelines.

Due to the presence of noise in the output content, a noise-filtering mechanism is designed. The process begins by filtering symbols and names of metal elements to capture items directly mentioning specific elements, as metals are often referred to in their elemental form. Next, it identifies alloy-related vocabularies by extracting a list of alloy terms from the model and performing full-name matching to accurately recognize alloys. Afterward, alphanumeric items are filtered based on common naming conventions of metal alloys, which frequently include numerical codes and letter prefixes or suffixes. Finally, for unfiltered words, the method identifies keywords directly associated with metals and alloys, such as "alloy" or "steel," classifying these items as metals. This approach is proven to be in line with our extraction expectations.

Data availability

The DOIs of articles on model construction and the complete source code for the models developed in this study are available at https://github.com/kun-ou-projects/tracking_progress_metallic_materials

Received: 8 October 2024; Accepted: 20 June 2025
Published online: 01 July 2025

References

1. Eswarappa Prameela, S. et al. Materials for extreme environments. *Nat. Rev. Mater.* **8**, 81–88. https://doi.org/10.1038/s41578-022-00496-z (2023).
2. Sun, F. et al. Comparative study on the stress corrosion cracking of X70 pipeline steel in simulated shallow and deep sea environments. *Mater. Sci. Eng. A* **685**, 145–153. https://doi.org/10.1016/j.msea.2016.12.118 (2017).
3. Vogl, T. et al. Radiation tolerance of two-dimensional material-based devices for space applications. *Nat. Comm.* **10**, 1202. https://doi.org/10.1038/s41467-019-09219-5 (2019).
4. David, N. V., Gao, X.-L. & Zheng, J. Q. Ballistic resistant body armor: Contemporary and prospective materials and related protection mechanisms. *Appl. Mech. Rev.* **62**, 050802. https://doi.org/10.1115/1.3124644 (2009).
5. Olabi, A. G. et al. A review on failure modes of wind turbine components. *Energies* **14**, 5241. https://doi.org/10.3390/en14175241 (2021).
6. Lee, C. et al. Strength can be controlled by edge dislocations in refractory high-entropy alloys. *Nat. Comm.* **12**, 5474 (2021).
7. El-Atwani, O. et al. Outstanding radiation resistance of tungsten-based high-entropy alloys. *Sci. Adv.* **5**, eaav2002. https://doi.org/10.1126/sciadv.aav2002 (2019).
8. El Atwani, O. et al. A quinary WTaCrVHf nanocrystalline refractory high-entropy alloy withholding extreme irradiation environments. *Nat. Comm.* **14**, 2516. https://doi.org/10.1038/s41467-023-38000-y (2023).
9. Wang, X., Huang, L., Xu, S. & Lu, K. How does a generative large language model perform on domain-specific information extraction? – A comparison between GPT-4 and a rule-based method on band gap extraction. *J. Chem. Info. Model.* **64**, 7895–7904 (2024).
10. Mikolov, T., Chen, K., Corrado, G. & Dean, J. Efficient estimation of word representations in vector space. *Proceedings of Workshop at ICLR* **2013** (2013).
11. Pennington, J., Socher, R. & Manning, C. GloVe: Global vectors for word representation. In Moschitti, A., Pang, B. & Daelemans, W. (eds.) *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543, https://doi.org/10.3115/v1/D14-1162(Association for Computational Linguistics, Doha, Qatar, 2014).
12. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C. & Solorio, T. (eds.) *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186, https://doi.org/10.18653/v1/N19-1423(Association for Computational Linguistics, Minneapolis, Minnesota, 2019).
13. Radford, A. & Narasimhan, K. Improving language understanding by generative pre-training (2018).
14. Tshitoyan, V. et al. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* **571**, 95–98 (2019).
15. Zhang, L. & He, M. Unsupervised machine learning for solar cell materials from the literature. *J. Appl. Phys.* **131**, 064902 (2022).

16. Huang, S. & Cole, J. M. Batterybert: A pretrained language model for battery database enhancement. *J. Chem. Infor. Modeling* **62**, 6365–6377 (2022).
17. Pei, Z., Yin, J., Liaw, P. K. & Raabe, D. Toward the design of ultrahigh-entropy alloys via mining six million texts. *Nat. Comm.* **14**, 54 (2023).
18. Aitchison, J. *Language Change: Progress Or Decay?* Cambridge Approaches to Linguistics (Cambridge University Press, 2001).
19. Hamilton, W. L., Leskovec, J. & Jurafsky, D. Diachronic word embeddings reveal statistical laws of semantic change. In Erk, K. & Smith, N. A. (eds.) *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1489–1501, https://doi.org/10.18653/v1/P16-1141(Association for Computational Linguistics, Berlin, Germany, 2016).
20. Lee, J. et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **36**, 1234–1240 (2020).
21. Blei, D. M., Ng, A. Y. & Jordan, M. I. Latent dirichlet allocation. *Journal of machine Learning research* **3**, 993–1022 (2003).
22. Yeh, J.-W. et al. Nanostructured high-entropy alloys with multiple principal elements: novel alloy design concepts and outcomes. *Adv. Eng. Mater.* **6**, 299–303 (2004).
23. Cantor, B., Chang, I. T. H., Knight, P. & Vincent, A. J. B. Microstructural development in equiatomic multicomponent alloys. *Mater. Sci. Eng. A* **375–377**, 213–218. https://doi.org/10.1016/j.msea.2003.10.257 (2004).
24. Gludovatz, B. et al. A fracture-resistant high-entropy alloy for cryogenic applications. *Science* **345**, 1153–1158 (2014).
25. George, E. P., Raabe, D. & Ritchie, R. O. *High-entropy alloys. Nat. Rev. Mater.* **4**, 515–534 (2019).
26. Youssef, K. M., Zaddach, A. J., Niu, C., Irving, D. L. & Koch, C. C. A novel low-density, high-hardness, high-entropy alloy with close-packed single-phase nanocrystalline structures. *Mater. Res. Lett.* **3**, 95–99 (2015).
27. Miracle, D. B. & Senkov, O. N. A critical review of high entropy alloys and related concepts. *Acta Mater.* **122**, 448–511 (2017).
28. Sinha, M. K. et al. Applications of sustainable techniques in machinability improvement of superalloys: a comprehensive review. *International Journal on Interactive Design and Manufacturing* **17**, 473–498 (2023).
29. Li, X., Du, D., Xia, Q. & Li, T. Mapping the structure and dynamics of global high-tech aerospace trade. *Global Networks* **24**, e12475 (2024).
30. Pikalova, E. Y., Kalinina, E. G., Pikalova, N. S. & Filonova, E. A. High-entropy materials in SOFC technology: Theoretical foundations for their creation, features of synthesis, and recent achievements. *Materials* **15**, 8783 (2022).
31. Senkov, O., Wilks, G., Miracle, D., Chuang, C. & Liaw, P. Refractory high-entropy alloys. *Intermetallics* **18**, 1758–1765. https://doi.org/10.1016/j.intermet.2010.05.014 (2010).
32. Geetha, M., Singh, A., Gogia, A. & Asokamani, R. Effect of thermomechanical processing on evolution of various phases in Ti-Nb-Zr alloys. *Journal of Alloys and Compounds* **384**, 131–144. https://doi.org/10.1016/j.jallcom.2004.04.113 (2004).
33. Andersson, J.-O., Helander, T., Höglund, L., Shi, P. & Sundman, B. Thermo-Calc & DICTRA, computational tools for materials science. *Calphad* **26**, 273–312. https://doi.org/10.1016/S0364-5916(02)00037-8 (2002).
34. Li, S. et al. Corrosion behavior of Sn-based lead-free solder alloys: a review. *J. Mater. Sci. Mater. Electron.* **31**, 9076–9090. https://doi.org/10.1007/s10854-020-03540-2 (2020).
35. Wee, S. et al. Review on mechanical thermal properties of superalloys and thermal barrier coating used in gas turbines. *Appl. Sci.* **10**, 5476. https://doi.org/10.3390/app10165476 (2020).
36. Ren, M. et al. Decarbonizing China's iron and steel industry from the supply and demand sides for carbon neutrality. *Appl. Energy* **298**, 117209. https://doi.org/10.1016/j.apenergy.2021.117209 (2021).
37. Yang, Q., Zhang, Y., Dai, W. & Pan, S. J. *Transfer Learning* (Cambridge University Press, 2020).
38. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. & Dean, J. Distributed representations of words and phrases and their compositionality. In *Neural Information Processing Systems* (2013).

Acknowledgements

We thank Mr. Mohammad Younes Araghi for Thermo-Calc calculations. We acknowledge the support of the Vice President for Research and Partnerships of the University of Oklahoma (OU) and the Data Institute for Societal Challenges. AR and SX are grateful for the startup funds provided by OU. YS was supported in part through the Utah NASA Space Grant Consortium, NASA Grant #80NSSC20M0103. SX was supported by a grant from the Research Council of the OU Norman Campus. Some of the computing for this project was performed at the OU Supercomputing Center for Education & Research (OSCER) at OU. Financial support was provided by the OU Libraries 'Open Access Fund.'

Author contributions

K.L. and S.X. conceived the study. K.L. designed the method. X.W. collected data and carried out modeling and analysis. A.R. contributed to the data collection and preparation. S.X. and Y.S. contributed to the interpretation of results. K.L., S.X. and Y.S. provided constructive suggestions on model optimization and provided critical feedback on the manuscript. All authors reviewed the manuscript.

Declarations

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/s41598-025-08356-w.

Correspondence and requests for materials should be addressed to K.L.

Reprints and permissions information is available at www.nature.com/reprints.

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