Computational Materials Science 230 (2023) 112441

Contents lists available at ScienceDirect

Computational Materials Science

journal homepage: www.elsevier.com/locate/commatsci

Full length article

A literature-mining method of integrating text and table extraction for materials science publications

Rui Zhang [a,b], Jiawang Zhang [a], Qiaochuan Chen [a], Bing Wang [a], Yi Liu [c,b], Quan Qian [a,b,c,d], Deng Pan [b,c], Jinhua Xia [a], Yinggang Wang [a], Yuexing Han [a,b,d,*]

[a] *School of Computer Engineering and Science, Shanghai University, 99 Shangda Road, Shanghai, 200444, China*
[b] *Zhejiang Laboratory, Hangzhou, 311100, Zhejiang, China*
[c] *Materials Genome Institute, Shanghai University, 333 Nanchen Road, Shanghai, 200444, China*
[d] *Key Laboratory of Silicate Cultural Relics Conservation (Shanghai University), Ministry of Education, 99 Shangda Road, Shanghai, 200444, China*

ARTICLE INFO

Dataset link: Materials-Science-Literature-Mining, https://github.com/han-yuexing/Materials-Science-Literature-Mining, https://github.com/han-yuexing/NER-SciBERT-Fasttext-BiLSTM-CRF

Keywords:
Literature mining
Named entity recognition
Table extraction
Material property prediction

ABSTRACT

Scientific literature, as an important tool to present research results, contains highly valuable information. Thus, there is an urgent need for mining methods to obtain key information from unstructured literature. However, existing literature mining efforts often ignore non-textual components which contain more detailed key information, such as tables. In this study, we propose a method for information processing to extract and analyze textual and tabular information from the large-scale literature of materials science. First, we propose a SciBERT-Fasttext-BiLSTM-CRF (SFBC) model for Named Entity Recognition (NER) in materials science literature, which combines generic dynamic word vectors (GDWVs) with domain-specific static word vectors (DSWVs). Second, a method is presented to extract material names, units, and compositions from the tables in literature. Compared to other table recognition methods, the method excels at extracting structured material composition information. Furthermore, Gradient Boosting Decision Tree algorithm is used to predict property trends including corrosion resistance, ductility, strength, and hardness on the basis of the material compositions, methods, properties and their changes extracted from texts and tables. The proposed method can be applied to predict material properties. Finally, we use stainless steel as an experiment example to validate our method. From 11,058 scientific papers on stainless steel, 2.36 million material entities and 7,970 material compositions are extracted. The extraction results are filtered and applied to predict four property trends of stainless steel. The method proposed in this paper can improve the accuracy of large-scale data extraction from material science literature, and the results can be used to guide the optimization of material properties and accelerate the pace of data-driven material design.

1. Introduction

Materials are an important cornerstone of human societal development. Metallic materials include iron, copper, aluminum, magnesium, zinc, etc. and their alloys. Inorganic non-metallic materials include ceramics, cement, glass, clay, etc. Organic polymer materials include plastics, rubber, cellulose, etc. Advanced composite materials refer to materials composed of two or more materials, such as carbon fiber composites, fiberglass, aluminum-based composites, etc. With the development of high-tech fields such as aerospace and biomedicine, higher demands have been placed on material performance and properties, leading to the emergence of new materials technology. New material technology refers to the interdisciplinary field of science and technology that focuses on the development, research, and application of novel materials. This field involves the systematic study of the theoretical principles, synthesis methods, and technological applications of new materials to meet the ever-increasing demands for advanced material properties and functions. It is characterized by multidisciplinary intersection, multi-level design, big data drive and so on. New materials technology has become one of the most important and promising fields in the 21st century [1]. In the Materials Genome Initiative, data mining, machine learning, and other technologies form the basis for big data-driven materials research methods, which are considered the fourth paradigm of materials research, unifying the other three paradigms in experimental, theoretical, and computational simulation aspects [2]. These methods have been applied in new material design [3–5], material property prediction [6–8], and other fields. The acquisition of high-quality data is the foundation and key of big data-driven new

[*] Corresponding author at: School of Computer Engineering and Science, Shanghai University, 99 Shangda Road, Shanghai, 200444, China.
E-mail address: han_yx@i.shu.edu.cn (Y. Han).

https://doi.org/10.1016/j.commatsci.2023.112441
Received 29 October 2022; Received in revised form 30 July 2023; Accepted 11 August 2023
Available online 31 August 2023

material research, and materials science literature contains a wealth of material knowledge and data. Using ‘Metal Material’ as a search keyword and ‘2017–2021’ as a time condition in the Elsevier ScienceDirect database, 630,000 scientific papers can be retrieved. Most of these scientific papers are stored and published as unstructured Portable Document Format (PDF) and contain a large number of textual components (e.g. abstracts and texts) and non-textual components (e.g. tables). It is crucial to quickly extract information and data from literature for the development of new materials and the discovery of new knowledge.

With the rapid development of natural language processing, researchers have proposed many tools and methods to extract and analyze information from texts on a large scale [9]. To explore the value of text mining results in scientific literature, Westergaard et al. conducted text mining on the full text of 15 million scientific papers of various types [10]. The difference between full-text mining results and abstracts is assessed by means of quantitative benchmarking, which demonstrates the effectiveness of scientific literature text mining for knowledge management and discovery. In the field of materials literature mining, Weston et al. mined material names, properties, characterization methods, phase descriptors, synthesis methods, and applications from the abstracts of 3.27 million materials science papers [9]. The mining results include more than 80 million materials science entities which a comprehensive materials science search and discovery engine constructed [11]. The automated tool MatScIE [12] was developed by Souradip et al. to extract relevant information from materials science literature for creating structured data which is used to simulate the materials. Kuniyoshi et al. proposed a framework integrating a text sequence annotation model and a numerical normalization module to analyze inorganic material trends [13]. And existing literature text information extraction work has not yet mined textual content together with non-textual information. However, non-textual information is also an important part of scientific literature, which contains important knowledge to strongly support textual content. Therefore, the existing literature text mining work has certain limitations.

Table, as a non-textual component embedded in the literature, is a class of important medium for conveying key information. Thus, table extraction methods have been rapidly developed. Chandran et al. designed a system for parsing tables in a tree structure based on heuristic rules, which takes horizontal and vertical table lines as clues [14]. Hao et al. proposed a deep learning method to detect table, which uses a table detection algorithm that incorporates heuristic rules and PDF document meta-information [15]. The SPLERGE [16] was proposed by Tensmeyer et al. for table structure recognition, which consists of a segmentation model with projection pooling and a merging model with grid pooling. The post-processing part parses the segmentation space generated by the row separator segmentation model to obtain the final table recognition results. Although the existing methods excel in identifying table contents, they do not adequately leverage the integration of table recognition results with relevant information, which limit the potential for achieving more in-depth information mining. In particular, the table information in scientific literature has not been fully exploited with textual information.

Machine learning has brought new opportunities to the field of material science. Machine learning uses existing data to train models and gives them the ability to predict unknown conditions. Therefore, machine learning plays a great role in computational materials science and is widely used in many fields such as materials analysis and materials design [17]. Some existing work [18–20] used machine learning to study the effects of chemical composition and process on glass properties. The trained predictive models can assist in the development of functional and bio-glass. Xiong et al. used a symbolic regression algorithm for feature screening on the NIMS steel dataset, and predicted fatigue strength, tensile strength, fracture strength and hardness on the feature set using the random forest algorithm [21]. In addition, a symbolic regression algorithm was used to calculate material property values. However, the existing work on studying material properties using machine learning algorithms suffers from narrow data sources, difficult data collection, and even manual methods that require a lot of time and human resources.

In this study, we propose a method for information extracting and applying in the large-scale materials science literature based on natural language processing, non-learning method, and machine learning. The proposed method includes a material text-oriented Named Entity Recognition (NER) model, a material composition table extraction method, and material property prediction. First, the NER model, which combined different word vector features, is trained in a corpus of 250 materials science papers. Thirteen entity features, such as material name, research aspect, technology, property, experiment condition and involved element, are selected for multidimensional statistical analysis. Second, a method is designed for table recognition and composition extraction, which is based on the structural characteristics of material composition tables detected from scientific literature in PDF format. The information similarity score of this method, in comparison to the recognition results of the PaddleOCR system, demonstrates a notable value of 93.59%. The material composition information in the tables, such as material names, elements, contents, units, etc., is identified and extracted. Then, material compositions, technologies, properties and their changes are filtered from the text and table extraction results. Using these data, the Gradient Boosting Decision Tree (GBDT) algorithm is used to train property change prediction models. Finally, 11,058 unannotated scientific papers on stainless steel are processed. Based on the results of the processing, the research hotspots and trends in the past 10 years are counted, and the relationships between entities are analyzed. In addition, property prediction models for corrosion resistance, ductility, strength and hardness are trained. The research contributions of our work are as follows:

- We propose a NER model called SFBC for combining generic dynamic word vectors (GDWVs) with domain-specific static word vectors (DSWVs) for material texts to extract material entity information accurately.
- We introduce a table recognition and composition extraction method for image-based material composition tables to obtain composition information from tables in scientific literature.
- We use GBDT algorithm to mine the potential relationships among material compositions, technologies and properties, which are obtained from material scientific literature, and predict the trend of material properties.

The following sections of paper are as follows: In Section 2, we define the category of named entities, display the manually constructed NER dataset, and present the acquisition and preprocessing methods of scientific literature. In Section 3, we introduce the proposed method of scientific literature information extraction and application, which is one of the deliverables of this work, including a NER model for extracting entities from literature text, a method for mining material composition from literature tables, and a machine learning-based property prediction method. In Section 4, we present experiments on entity extraction of scientific literature text using different NER models on two NER datasets. Experiment on material composition table extraction and evaluation method are also presented. In Section 5, we demonstrate the application of the methodology proposed in this paper to 11,058 scientific papers on stainless steel, including data statistical analysis and property prediction based on data, which is another deliverable of this paper. Finally, the conclusion and future works are described in Section 6.

2. Data preparation

To extract material entities from the texts of the literature, we manually create a stainless steel named entity recognition dataset for training and validating the SFBC model. In addition, to apply the literature mining method proposed in this paper, including text mining and table mining, we also collect 11,058 scientific papers on stainless steel.

2.1. Entity category definition

We define the mining objects of the named entity recognition model SFBC using 13 entity labels, including: material name, research aspect, technology, method, material property, property value, experiment name, experiment condition, condition value, experiment output, equipment used, involved element, and applicable scenario. Each entity specifically defined as shown below:

Material Name (MN): In various texts within the literature, we can find a multitude of sentences where different material names are mentioned, encompassing either their complete names or corresponding abbreviations, such as 'Fe and Cr phase separation in ferrite, causing 475 °C-embrittlement, was studied after very short aging times in super duplex stainless steel and hyper duplex stainless steel plates and welds' [22].

Research Aspect (RA): Research aspect indicates the focus of discussion and research in the scientific literature, such as 'Moreover, from the analysis of scanning strategy, the SLM sample with 90° rotation shows the best corrosion resistance followed by the sample with 0° rotation, and the sample with 67.5° rotation shows the worst corrosion resistance' [23].

Technology (Tech): Technology indicates the processes involved in the scientific literature, such as 'Some specimens were cold-rolled again to investigate the work-hardening behavior(cold-rolled specimens)' [24].

Method (Me): Method indicates the research technique used in the material research experiment, such as 'The microstructures were analyzed by electron backscatter diffraction (EBSD) with a field-emission scanning electron microscope' [25].

Property (Prop): Material property represent the characteristic properties of the material studied in the scientific literature, such as 'Duplex stainless steels (DSS) are applied widely in offshore engineering, petrochemical and nuclear power fields because the two-phase structure of ferrite and austenite possesses good toughness, high strength and excellent corrosion resistance' [26].

Property Value (PV): Property value indicate trends or specific values of the studied properties of the material, such as 'Duplex stainless steels (DSS) are applied widely in offshore engineering, petrochemical and nuclear power fields because the two-phase structure of ferrite and austenite possesses good toughness, high strength and excellent corrosion resistance.'[26] and 'The strengthening effect caused by the microstructures resulted in a tensile strength of 627MPa with a maximum elongation of 50%' [27].

Experiment Name (EN): Experiment name indicates that the scientific literature specifies the experiment performed during the study, such as 'The tensile experiment was conducted using an MTS testing machine in accordance with ISO 6892, and the fracture of the sample was analyzed using SEM' [28].

Experiment Condition (EC): Experiment condition indicates the influencing factors, variables, etc. involved in the experiment, such as 'The EIS measurement was performed at 25 °C in 3.5wt%-NaCl solution over a frequency range of 10mHz to 100 kHz, with an acquisition rate of 10 points per decade, with a signal amplitude of 10 mV at the open circuit potential (OCP)' [28].

Condition Value (CV): Condition value indicates the specific value of the experimental condition, such as 'The EIS measurement was performed at 25 °C in 3.5wt%-NaCl solution over a frequency range of 10 Hz to 100 kHz, with an acquisition rate of 10 points per decade, with a signal amplitude of 10mV at the open circuit potential (OCP)' [29].

Experiment Output (EO): Experiment output indicates the results obtained from the experiment, such as 'The flow stress curves under different conditions were depicted in Figure 6' [30].

Equipment Used (EU): Equipment used indicates the name of the experimental equipment used in various material experiments, such as 'The solidus and liquidus temperatures of each SS-B4C were determined using a differential scanning calorimeter' [31].

Involved Element (IE): Involved element indicate elements highlighted by researchers during the study, which may be components of the material itself or may be added additionally by researchers during the experiment, such as 'Yttrium, as a reactive element, has similar features with cerium and laudanum' [32].

Applicable Scenario (AS): Applicable scenario refers to the range or circumstances in which a material can exhibit its inherent characteristics and functionalities in specific environments or application conditions, such as 'These results are helpful to promote the application of duplex stainless steel in the fields of nuclear power fields' [26].

2.2. NER dataset preparation

To create our NER dataset, 250 English-language stainless steel papers are collected from the Elsevier ScienceDirect database, which are publicly available. These papers are retrieved by using the keyword 'stainless steel' and setting the time range from 2012 to 2021, with the results sorted by relevance in descending order. 25 papers are collected from the search results of each year, obtaining a total of 250 papers for annotation. We conduct processing on 250 scientific papers, from which we extract 2,453 sentences, and apply sequence labeling to these sentences using the Doccano [33] tool. The approximate steps for using the Doccano tool for data annotation are as follows: installation of the tool, creation of annotation projects, data uploading, manual annotation, and data export. Prior to manual annotation, we create thirteen material entity categories in the Doccano, named as: material name, research aspect, technology, method, material property, property value, experiment name, experiment condition, condition value, experiment output, equipment used, involved element, and applicable scenario. Subsequently, the sentences to be annotated are uploaded to the doccano tool. A feasible approach is to place the sentences in a TXT file, with each sentence occupying a separate line. The TXT file is then uploaded to doccano. Finally, each word or consecutive words in each sentence are manually selected, and the corresponding material entity category is chosen to complete the named entity annotation.

Then, the annotated data are converted into BIO (B-Begin, I-Inside, O-Outside) format [34]. Take 'flow stress' as an example, the labeled category is 'Research Aspect'. After converting the BIO format, 'flow' is labeled as 'B-Research Aspect' and 'stress' is labeled as 'I-Research Aspect'. Take 'DIM' as an example, the labeled category is 'Material Name'. After conversion, 'DIM' is labeled as 'B-Material name'. Other words that are not labeled are all converted to 'O' category. The 2,453 annotated data are divided into training and test sets in an 8:2 ratio, and the SFBC model proposed in this paper is trained using the training set. The evaluation metrics of the model are calculated on the test set.

2.3. Literature mining dataset preparation

To apply our method in materials science, we collect 11,058 publicly available English-language scientific papers on stainless steel from the Arxiv and Elsevier ScienceDirect databases. The collected scientific literature is in PDF format and cannot be directly used for literature mining. In order to perform text mining, we preprocess the scientific literature in PDF format. The preprocessing steps included converting PDF to Word, extracting the textual content of the paper, and segmenting the text into sentences. In order to extract composition tables, we use the fitz [35] and PyMuPDF [35] tool libraries in Python to convert the paper in PDF to image. Then, the collected materials science literature can be utilized for text mining and extraction of compositional tables. We utilize the SFBC model for text mining and extract 13 types of material entities from the text in literature. In addition, we apply table recognition and composition extraction method to obtain information about material compositions from tables in image format.

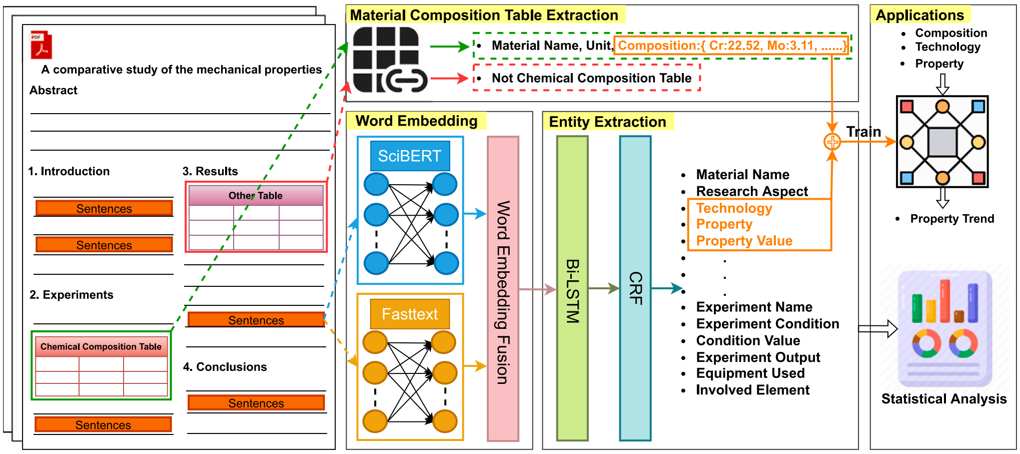


Fig. 1. Information extraction and application framework for literature of materials science. 'Material Composition Table Extraction' module can extract material names, elements, contents and units. 'Word Embedding' and 'Entity Extraction' modules are combined into SFBC model for material texts. 'Applications' module can statistically analyze the results of literature mining and predict the property trend.

3. Methodology

The proposed method for extracting and applying large-scale scientific literature information is shown in Fig. 1, including material text information extraction, material composition table extraction, and material property prediction. Text information extraction uses a sequence annotation method to extract target entities from text in the scientific literature. Material composition table extraction uses morphological image processing methods to extract material names, units, and compositions from tables in the scientific literature. Material property prediction uses the text and table extraction results to predict the trends of corrosion resistance, ductility, strength, and hardness using machine learning algorithm.

3.1. Material text NER model

Our work employs natural language processing techniques to perform data mining on literature text, and proposes a named entity recognition (NER) model called SciBERT-Fasttext-BiLSTM-CRF (SFBC). The goal of the SFBC model is to extract 13 types of material entities from the text of materials science literature. It takes sentences from the literature as input to the model and outputs the entity type of each word in the sentence. The structure of SFBC is shown in Fig. 2. Due to the peculiarities in expressing sentences in material science literature, the SFBC model combines the dynamic word vectors from a general-domain language model with the static word vectors from a material-domain language model, which enriches each word vector with both contextual information and material-domain knowledge. This approach compensates for the lack of material-domain knowledge in general-domain dynamic language model and solves the problem of lack of contextual information in material-domain static language model during representation. The SFBC model consists of a general-domain dynamic language model SciBERT [36], a material-domain static language model Fasttext [37], and a word vector fusion module. The downstream network structure includes a Bi-directional Long Short-Term Memory (Bi-LSTM), a Fully Connected layer (FC), and a Conditional Random Field (CRF).

The SciBERT [36] is based on the BERT [38] and is fine-tuned on a dataset of 1.14 million scientific papers. The vocabulary of SciBERT consists of 31,090 words or subword units. In the field of natural language processing, 'token' can be understood as a lexical unit that refers to a word or its subcomponent. Before vectorized with SciBERT, the input sentence needs to be preprocessed and converted into several 'token' with 'BertTokenizer', as shown in Fig. 2. The 'BertTokenizer' is a class in HuggingFace [39] and is initialized with the weight of SciBERT. For instance, when representing the sentence 'electropolishing of 316 stainless steel' using the SciBERT, 'of', '316', 'stainless' and 'steel' are in the SciBERT's vocabulary and are considered as known 'token', while 'electropolishing' is not in the vocabulary and is considered an unknown 'token'. The 'electropolishing' is segmented into 'electro', '##pol', and '##ishing', where all three 'token' are present in the vocabulary, and this segmentation strategy is known as 'WordPiece'[40]. Eventually, through the vectorization of SciBERT, all 'token' in the input sentence will be represented as generic dynamic word vectors (GDWVs). There are many domain-specific specialized words in the materials science literature, and most of these words are not in the SciBERT's vocabulary. Therefore, when the SciBERT [36] is directly used for NER of material literature, its performance will be limited.

Kim et al. obtained the Fasttext [37] for materials synthesis based on the native Fasttext [41] which is trained on 2.5 million materials science papers. Here, we use the Fasttext [37] of Kim to obtain domain-specific static word vectors (DSWVs) of words in material texts. Kim's Fasttext can effectively perform static characterization of materials science words. However, Fasttext does not integrate contextual information when characterizing words, and it has limitations in natural language sequence annotation tasks.

Therefore, the proposed method combines the GDWVs and material DSWVs for the fused word vectors used in the material domain NER. To compensate the lack of domain-specific features for SciBERT [36], the material domain Fasttext [37] is introduced to complement its features.

Use the word sequence to represent the input sentence, and the word sequence $Seq = [W_1, W_2, \dots, W_i]$ is input into SFBC, where W_i is the word in the input sentence and i is the position of the word in the sentence. Thus, the word W_i in the sequence can be cut into 'token' E_{ik_i} by 'BertTokenizer', denoted as $W_i = [E_{i1}, E_{i2}, \dots, E_{ik_i}], k_i \geq 1$. k_i is the number of word W_i divided into 'token'. Then, E_{ik_i} is represented as an EV_{ik_i} vector with SciBERT [36]. Thus, the dynamic word vector for the W_i is obtained as $BV_i = [EV_{i1}, EV_{i2}, \dots, EV_{ik_i}]$, $k_i \geq 1$. EV_i is a 768-dimensional vector obtained by vectorizing the 'token' using SciBERT. BV_i is a dynamic vector representation of the word W_i in the context of the Seq. The word W_i in the sequence Seq is statically characterized by the material-domain Fasttext [37] to obtain FV_i. FV_i is a static vector representation of the word W_i in the context of the Seq, and it is a 100-dimensional vector. Combine BV_i and FV_i to get the fusion word vector WV_i of the word W_i, which is a vector of $k_i * 868$ dimensions and expressed as

$$WV_i = [\{EV_{i1}, FV_i\}, \{EV_{i2}, FV_i\}, \dots, \{EV_{ik_i}, FV_i\}], k_i \geq 1. \tag{1}$$

Finally, the input word sequence Seq is characterized as

$$SV = [WV_1, WV_2, \dots, WV_i], \tag{2}$$

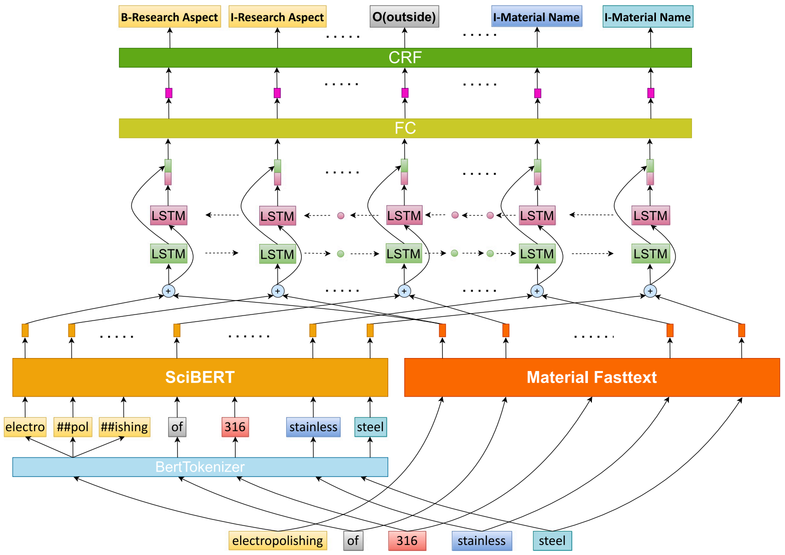


Fig. 2. Overview of SFBC which is a NER model for material texts. The SFBC model combines GDWVs and material DSWVs to extract text features. The downstream uses BiLSTM network, FC layer and CRF to finally predict the entity category of each word in the input sentence.

which is a vector of $(k_1 + k_2 + \cdots + k_i) * 868$ dimensions and input into the downstream BiLSTM-CRF network.

The Long Short-Term Memory (LSTM) [42] network is an improved model based on Recurrent Neural Networks (RNN) to solve the gradient disappearance and gradient explosion problems [43]. The BiLSTM can fuse forward LSTM and reverse LSTM features, to effectively learn the context information of text sequences. In addition, the BiLSTM layer can help the model better understand the semantic and syntactic information in the sentence, leading to more accurate entity recognition. After the processing of BiLSTM, the dimension of the vector remains unchanged, and the $(k_1 + k_2 + \cdots + k_i) * 868$ dimension vector is input to the FC layer.

The output of BiLSTM is a $[(k_1 + k_2 + \cdots + k_i) * 868]$-dimensional feature vector. The FC layer maps the 868-dimensional vector to the N-labeled category space, i.e., $\mathbb{R}^{868} \mapsto \mathbb{R}^N$. N is the number of entity label categories output in the SFBC, and it depends on the number of predefined entity categories in the dataset and the annotation strategy used for the dataset. If the predefined entity categories in the dataset are n and the annotation strategy used is BIO format [34], ‘B’ and ‘I’ are spliced in the head of the predefined entity names in the dataset, respectively. ‘B’ represents the beginning word of the entity name and ‘I’ represents the other part of the entity name. Thus n entity classes are transformed into $2 * n$ and all unlabeled words in the dataset are divided into ‘O’ classes, $N = 2 * n + 1$. In our dataset, $n = 13$, thus $N = 27$. Therefore, the FC layer transforms an $[(k_1 + k_2 + \cdots + k_i) * 868]$-dimensional vector to a $[(k_1 + k_2 + \cdots + k_i) * 27]$-dimensional vector.

The output layer of SFBC is CRF, which is a conditional probabilistic undirected graph model. The $[(k_1 + k_2 + \cdots + k_i) * 27]$-dimensional feature vector which is the output of FC layer and input into CRF. The CRF layer is responsible for modeling the dependencies between the output labels and uses transition matrix to calculate the conditional probability of the label sequence. Each ‘token’ position selects the most probable label from 27 label categories. Finally, the output consists of the entity label corresponding to each input ‘token’, and the length of the resulting sequence is $k_1 + k_2 + \cdots + k_i$.

In summary, our proposed NER method takes two types of language vectors as input and achieves material entity extraction from sentences. The entity category representation of SFBC model output is in BIO format [34]. As shown in Fig. 2, the sentence ‘electropolishing of 316 stainless steel’ is input into SFBC as an example. The ‘BertTokenizer’ slices the input sentence to [‘electro’, ‘##pol’, ‘##ishing’, ‘of’, ‘316’, ‘stainless’, ‘steel’], and SciBERT performs dynamic vector representation for each item. The Fasttext performs a static vector representation for each word in the input sentence. The dynamic and static word vectors are combined and input to the downstream network. The final CRF layer outputs results as [B-Research Aspect, I-Research Aspect, I-Research Aspect, O, B-Material Name, I-Material Name, I-Material Name].

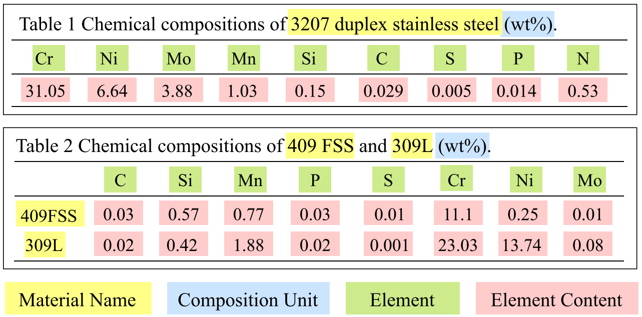


Table 1 Chemical compositions of 3207 duplex stainless steel (wt%).

Cr	Ni	Mo	Mn	Si	C	S	P	N
31.05	6.64	3.88	1.03	0.15	0.029	0.005	0.014	0.53

Table 2 Chemical compositions of 409 FSS and 309L (wt%).

	C	Si	Mn	P	S	Cr	Ni	Mo
409FSS	0.03	0.57	0.77	0.03	0.01	11.1	0.25	0.01
309L	0.02	0.42	1.88	0.02	0.001	23.03	13.74	0.08

Material Name
Composition Unit
Element
Element Content

Fig. 3. Composition table structure diagram. Material composition includes material name, unit, element, and element content. The number of materials in multi-material composition tables is no less than 2 and the title of multi-material composition tables does not necessarily contain the material name, but it must contain the unit.

3.2. *Material composition table extraction method*

The goal of our proposed table composition extraction method is to extract material composition information from tables in PDF format literature. Most scientific papers are published in PDF format, which makes the tables embedded in them inaccessible. Therefore, we apply YOLOv3 [44] to crop the table regions from the layout images of scientific papers. The proposed method processes image-based tables as input and extracts material composition information. The material compositions are outputted in JSON format, containing material name, unit, element, and element content. This method can extract both single-material composition tables and multi-material composition tables, as shown in Fig. 3. The structure of the method is shown in Fig. 4.

There are various forms of tables in scientific literature. In order to avoid the influence of different lines in table, we need to remove the table borders and enhance the text area features from the image-based table which is cut from the literature layout image. The process of using morphology-based methods to preprocess image-based table is shown

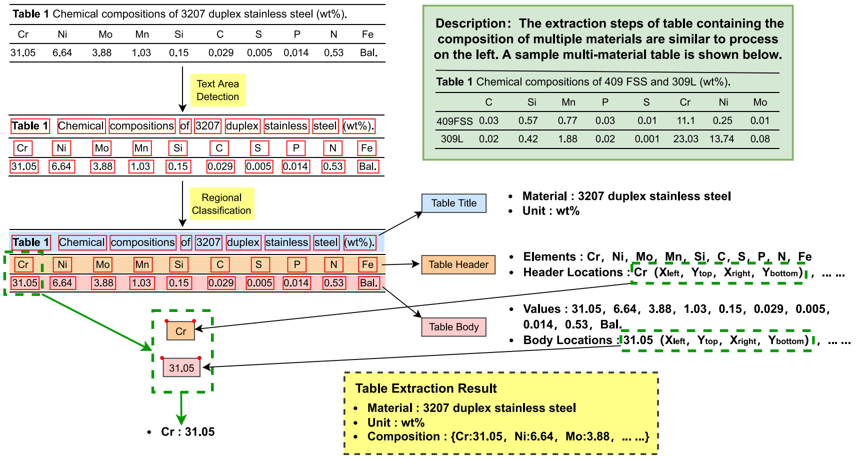


Table 1 Chemical compositions of 3207 duplex stainless steel (wt%).
Cr Ni Mo Mn Si C S P N Fe
31.05 6.64 3.88 1.03 0.15 0.029 0.005 0.014 0.53 Bal.
Text Area Detection
Regional Classification
Table Title
Table Header
Table Body
Material : 3207 duplex stainless steel
Unit : wt%
Elements : Cr, Ni, Mo, Mn, Si, C, S, P, N, Fe
Header Locations : Cr (Xleft, Ytop, Xright, Ybottom), ...
Values : 31.05, 6.64, 3.88, 1.03, 0.15, 0.029, 0.005, 0.014, 0.53, Bal.
Body Locations : 31.05 (Xleft, Ytop, Xright, Ybottom), ...
Cr
31.05
Cr : 31.05
Description: The extraction steps of table containing the composition of multiple materials are similar to process on the left. A sample multi-material table is shown below.
Table 1 Chemical compositions of 409 FSS and 309L (wt%).
C Si Mn P S Cr Ni Mo
409FSS 0.03 0.57 0.77 0.03 0.01 11.1 0.25 0.01
309L 0.02 0.42 1.88 0.02 0.001 23.03 13.74 0.08
Table Extraction Result
Material : 3207 duplex stainless steel
Unit : wt%
Composition : {Cr:31.05, Ni:6.64, Mo:3.88, ...}

Fig. 4. Framework of table recognition and composition extraction method. We convert scientific literature in PDF format into images. The target detection model is used to detect and cut images of table from the layout images of scientific literature. According to the title, the table is divided into material composition table and non-material composition table. For the material composition table, the technologies of text area detection, area classification and character block pairing are used to extract the material names, units and compositions.

in Fig. 5. In Fig. 5, the function of morphological dilation operation is to merge the background (black pixel area) that touches the foreground (white pixel area) into the foreground in a binary image, thereby achieving outward expansion of the foreground boundary points. The opposite operation is morphological erosion, which can achieve inward shrinkage of the foreground boundary. The essence of morphological open operation is to first perform morphological erosion on the binary image, and then perform morphological dilation. The basic function of this operation is to smooth the contour of the foreground area. We use morphological open operation and set different parameters to extract horizontal and vertical lines in the table separately. Using bitwise operation, the horizontal and vertical line images are combined with the original image to obtain a table image without any lines. Then, the table image without lines is processed using morphological dilation operation to form character block region from single character. The coordinates of the character block region are obtained using contour detection techniques based on binary images. Finally, we use the CRNN [45] model to recognize the text in the character block regions. The coordinate information of the character block region is detected as L_{kj}, and the text content of the character block region is recognized as T_{kj}, where k is the row position of the character block in the table and j is the serial number of the character block in the same row. L_{kj} is expressed as $L_{kj} = (X_l, Y_t, X_r, Y_b)$, where (X_l, Y_t) is the coordinates of the upper left corner of the character block region and (X_r, Y_b) is the coordinates of the lower right corner. Based on the span of Y_t and Y_b values in the region L_{kj}, the character blocks in the same line are calculated. We denote the kth line of text as TL_k as $TL_k = [T_{k1}, T_{k2}, \ldots, T_{kj}], j \geq 1$.

In order to extract material composition table information, it is necessary to filter the composition tables from scientific literature tables, and also break down composition tables into title, header and body. Table titles in scientific literature are usually above the table frame lines, so the table can be split according to the position of the top horizontal line. The upper part of the top horizontal line is the table title, and the lower part is the table header and body. Using regular expressions to match whether there are keywords ‘composition’ in the table title, to determine whether it is a composition table or not.

Counting the number of text lines in the table header and body, if the number is equal to 2, the first line is the header and the second line is the body, and the table is a single material composition table. When the number of text lines is greater than 2, using Word2Vec [46] language model in material domain, representing character block text T_{kj} in header and body parts as text vector V_{kj}, and then transforming text line TL_k into text line vector $VL_k = V_{k1} + V_{k2} + \cdots + V_{kj}$. Then calculate the cosine similarity S_k of the current text line vector VL_k and the next line vector VL_{k+1}.

$$S_k = \frac{VL_k \cdot VL_{k+1}}{\|VL_k\| \|VL_{k+1}\|}, \tag{3}$$

The higher the similarity of text content, the higher the cosine similarity score. The table body usually shows the same type of data, so there is a very high similarity between text lines in this area. Through experiments we found that the cosine similarity between text line vectors within the body is greater than 0.6. The similarity between the last text line of the header and the first text line of the body is low, and the similarity between multiple text lines within the header is also low. If there are two or more $S_k < 0.6$, it indicates that there are multiple text lines in the header, and this type of table is not extracted. If it is calculated that there is only one $S_k < 0.6$, it can be inferred that there is only one text line in the header and multiple text lines in the body, and this table is a multi-material composition table.

For composition tables containing a single material, the title includes the material name and unit, the header includes the elements, and the body includes the element contents. For composition tables containing multiple materials, the title includes the unit, the header includes the elements, and the body includes the material names and element contents. Using the SFBC model and regular expressions, material names and units can be extracted from the titles of single-material composition table. Using regular expressions, units can be extracted from the titles of multi-material composition table, and then material names can be obtained from the first character blocks of each row in the body part.

Based on the numerical span of X_l and X_r in the coordinate information L_{kj} of the character blocks in the header, find the character block $T_{(k+t)i}$ that is directly below the character block T_{kj}, where T_{kj} is an element and $T_{(k+t)i}$ is its corresponding content. $1 \leq t \leq n$ and n is the number of text lines in the table body and the number of material compositions in the table.

Furthermore, an evaluation of the extracted results from the composition table is required. After investigation, We have not found evaluation method for composition table extraction. Therefore, we propose a method that utilizes PaddleOCR [47] to assist in evaluating the extraction results of composition tables. The evaluation method is shown in Fig. 6.

PaddleOCR is an open source multilingual OCR framework, in which the PP-Structure structured document analysis toolkit supports

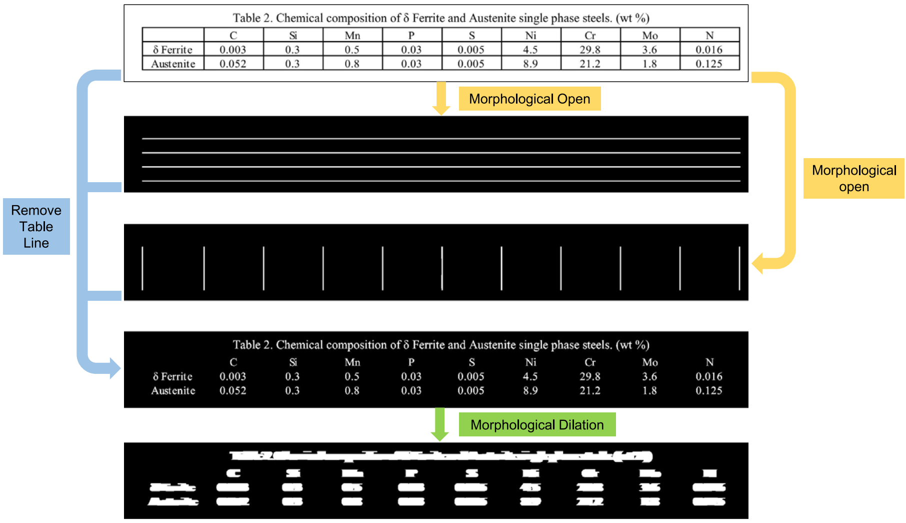


Table 2. Chemical composition of δ Ferrite and Austenite single phase steels. (wt %)

	C	Si	Mn	P	S	Ni	Cr	Mo	N
δ Ferrite	0.003	0.3	0.5	0.03	0.005	4.5	29.8	3.6	0.016
Austenite	0.052	0.3	0.8	0.03	0.005	8.9	21.2	1.8	0.125

Fig. 5. The flowchart of preprocessing for image-based tables. Using the morphological open operation, the table lines are obtained and further processed to remove them. In addition, the morphological dilation operation is used to expand the features of the text area for detection.

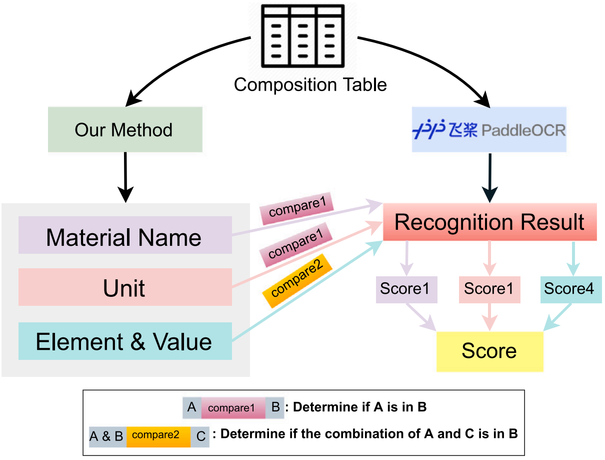


Fig. 6. Evaluation method of material composition extraction results. We calculate the material name, unit and composition extracted by our method with the PaddleOCR table recognition result respectively, and the three scores form the final similarity score.

table recognition. The PaddleOCR is validated on the PubTabNet [48] table recognition dataset using Tree-Edit-Distance-based Similarity (TEDS) [48] for table recognition results evaluation. The TEDS score of PaddleOCR is 93.32, so PaddleOCR has a good effect on table extraction. PaddleOCR table recognition result is set **PR** $= \{t_i\}_{i=1}^{n}$, where t_i is the text in the cell and n is the number of cells, and the table header area is also considered as a cell. Select the element e_j from the **PR**, obtain the element value c_j corresponding to the element e_j according to the cell coordinate information, and construct the composition set **PC** $= \{(e_j, c_j)\}_{j=1}^{m}$, where m the number of elements in the table. We denote the material name extracted by our method as M, the unit as U, and the composition set as **OC**. The calculation process of information similarity score is shown in Algorithm 1.

In summary, we propose a composition table extraction method based on morphological processing, which decomposes the structure of composition tables according to table lines, cell positions and text information, and can extract material names, units and composition information from titles, headers and bodies. Additionally, we also present an evaluation method specifically designed for assessing the extraction results of composition tables.

Algorithm 1 Table Recognition and Composition Extraction Similarity Score Algorithm

```
Require: $\mathbf{PR}, \mathbf{PC}, M, U, \mathbf{OC}$
Ensure: information similarity score
$curTotalScore \leftarrow$ current total score, initial value is 0;
$cellCount \leftarrow$ the number of objects to be extracted in the table, assigned as $|\mathbf{PC}|+2$;
if $M$ in PR then
    $curTotalScore \leftarrow curTotalScore + 1$;
else
    $matNameScore \leftarrow$ maximum value of semantic similarity between $M$ and each text in PR;
    $curTotalScore \leftarrow curTotalScore + matNameScore$;
end if
if $U$ in PR then
    $curTotalScore \leftarrow curTotalScore + 1$;
else
    $unitScore \leftarrow$ maximum value of semantic similarity between $U$ and each text in PR;
    $curTotalScore \leftarrow curTotalScore + unitScore$;
end if
$compScore \leftarrow |\mathbf{PR} \cap \mathbf{OC}|$;
$curTotalScore \leftarrow curTotalScore + compScore$;
return $curTotalScore/cellCount$;
```

3.3. Property trend prediction model

Gradient Boosting Decision Tree (GBDT) [49] algorithm is used to train property prediction models, which can predict the change trend of property. The trained models take material composition and technology as input, and output the change trend of property. The results of literature mining are used as training data to train models. Using the SFBC, 13 types of material entities can be mined from text in scientific literature. Material composition information can be obtained from tables using table recognition and composition extraction method. From the extraction results of texts and tables, material compositions and material entities are selected to train the property trend prediction models on GBDT, and the categories of material entities are technology, property and property value. In the mining results of the SFBC model, 77.3% of property value entities are presented in literature in the form of change trends, e.g., 'heat treatment was shown to decrease yield/tensile strength'. Therefore, the model we trained can only be used to predict property change trends, not specific values.

GBDT algorithm, as an important algorithm in integrated learning [50], has the advantages of preventing overfitting and strong generalization ability and is widely used in classification prediction [51]. GBDT is a classification and regression tree (CART) based model which uses the idea of boosting iterations. The original training set was used to obtain the first decision tree. Then, the goal in each iteration thereafter is to fit the residuals of the previous round of weak learners [52]. The modeling process of GBDT aims at minimizing the squared loss function of the current learner. Multiple iterations are performed to reduce the training residuals, and finally the results of all trained regression trees are summed to obtain the final prediction results [53]. Material composition, technology and property are input into the GBDT prediction model, and the output is the trend of property.

We employ material named entity recognition and composition table extraction method to extract information from 11,058 stainless steel literature papers. From the extraction results of text and table, we collect 376 items (composition, technology, corrosion resistance trend) data for training the corrosion resistance prediction model, 313 items (composition, technology, ductility trend) data for training the ductility prediction model, 265 items (composition, technology, hardness trend) data for training the hardness prediction model, and 756 items (composition, technology, strength trend) data for training the strength prediction model. The technology and property trend data are derived from texts and the material composition data from tables. Material compositions include chromium, nickel, molybdenum, nitrogen, manganese, aluminum, silicon, titanium, copper, carbon and cobalt elements. Technologies include annealing, heat treatment, hot rolling, cold rolling, quenching, tempering, selective laser melting, and laser powder bed fusion. The trend of property change is divided into positive increase and negative decrease.

In summary, we use GBDT to train property prediction models on corrosion resistance, ductility, strength and hardness data respectively. The trained models can predict the change trend of properties based on material composition and technology.

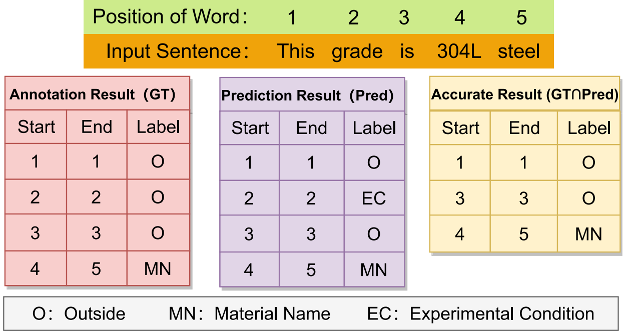
4. Experiments

Our operating environment is Intel(R) Xeon(R) Silver 4210R CPU @2.40 GHz and GeForce RTX 3090 TURBO 24G in Python 3.8.0 and Torch 1.7.1.

4.1. Material text NER

The SFBC model proposed in this paper for scientific literature in the materials domain combines the GDWVs and material DSWVs for feature fusion, and 13 kinds of entity information are extracted from literature of materials science.

We combine two dynamic language models, BERT [38] and SciBERT [36], with two static language models in the material domain, Word2Vec [46] and Fasttext [37], to obtain fusion language models, including BERT+Word2Vec, BERT+Fasttext, SciBERT+Word2Vec and SciBERT+Fasttext. The four fusion language models are compared with other single language models, including Word2Vec [46], Mat2Vec [54], Fasttext [37], BERT, ALBERT [55], SciBERT, ClinicalBERT [56], BioBERT [57], MatBERT [58], MatSciBERT [59] and MatTPUSciBERT [60]. Among these models, Word2vec is trained on 640,000 materials papers; Mat2Vec is trained on 3.3 million abstracts of materials papers; Fasttext is trained on 2.5 million materials papers; MatBERT is trained on 2 million materials papers; MatSciBERT is fine-tuned on SciBERT with 150,000 materials papers, and MatTPUSciBERT is fine-tuned on SciBERT with 700,000 materials papers using Tensor Processing Unit (TPU). In the experiments, the downstream networks of all NER models are BiLSTM-CRF. To construct the dataset, 2,453 sentences are collected from 250 stainless steel papers and performed sequence tagging on them. Divide the dataset into a training set (1,956 items) and a test set (467 items) according to 8:2. The distribution of the number of 13 entity label categories in the training and test sets also roughly matched the 8:2 ratio, as shown in Table 1. All experimental parameters are set to 'epoch'=500, 'batch size'=32, 'learning rate'=0.002 and 'dropout'=0.4.



Position of Word:	1	2	3	4	5
Input Sentence:	This	grade	is	304L	steel

Annotation Result (GT)

Start	End	Label
1	1	O
2	2	O
3	3	O
4	5	MN

Prediction Result (Pred)

Start	End	Label
1	1	O
2	2	EC
3	3	O
4	5	MN

Accurate Result (GT∩Pred)

Start	End	Label
1	1	O
3	3	O
4	5	MN

O: Outside MN: Material Name EC: Experimental Condition

Fig. 7. A schematic diagram of evaluation metrics for named entity recognition method. The manual annotation results of the sample sentence are in the GT table, the prediction results of the NER model for this sentence are in the $Pred$ table, and the accurate prediction results are in the $GT \cap Pred$ table. 'Start' and 'End' indicate the start and end positions of entities in input sentence, and 'Label' is entity type.

Based on the NER evaluation metrics [12], the annotation of the sentence is called ground truth (set GT) and the prediction result is called the predicted token (set $Pred$). As shown in Fig. 7, the manual annotation results of the sample sentence are in the GT table, the prediction results of the NER model for this sentence are in the $Pred$ table, and the accurate prediction results are in the $GT \cap Pred$ table. 'Start' and 'End' indicate the start and end positions of entities in input sentence, and 'Label' is entity type. The Precision (P_{ner}), Recall (R_{ner}) and F1-score ($F1_{ner}$) are calculated as shown in Eqs. (4)–(6), respectively. The comparison of the total scores of Precision, Recall, F1-score, Average Value (AVG), and Standard Deviation (SD) for different models on the test set is shown in Table 2, and the comparison of the F1-score of 13 entity label categories for different models on the test set is shown in Table 3.

$$P_{ner} = \frac{|GT \cap Pred|}{|Pred|} \tag{4}$$

$$R_{ner} = \frac{|GT \cap Pred|}{|GT|} \tag{5}$$

$$F1_{ner} = \frac{2P_{ner}R_{ner}}{P_{ner} + R_{ner}} \tag{6}$$

The comparison reveals that the strategy of combining the GDWVs and DSWVs can maintain the utterance sequence contextual information while fusing material domain-specific lexical features, effectively improving the performance of NER. Among four fusion models, SciBERT+Fasttext-BiLSTM-CRF is the best one, its $F1$ score of 80.08, which is 2.46 higher than the $F1$ score of the baseline model BERT-BiLSTM-CRF and 0.14 lower than the $F1$ score of MatSciBERT-BiLSTM-CRF. Compared with MatSciBERT, which is fine-tuned on 150,000 scientific papers based on SciBERT and requires a large amount of computational resources, our method can achieve a similar performance by fusing SciBERT and Fasttext word embeddings without any further training on data. In addition, the $F1$ scores of BERT+Word2Vec, BERT+Fasttext and SciBERT+Word2Vec also showed significant improvements compared to Word2Vec, Fasttext, BERT and SciBERT.

In addition, Weston et al. [9] manually annotated NER for 800 abstracts of materials literature, and referred to this dataset as MatData in this paper. MatData materials NER dataset defines seven categories of materials entities: inorganic material (MAT), synthesis method (SMT), characterization method (CMT), material application (APL), material

Table 1
The number distribution of the thirteen entity labels in training set and test set. We collect and annotate 2453 sentences from 250 stainless steel papers. The distribution ratio of each entity label category in the train set and verification set is about 8:2. Overall, there are 4468 entity tags in the train set and 1127 entity tags in the verification set, and the total distribution ratio also conforms to the 8:2 rule.

Entity label	Train set	Test set	Ratio
Material Name(MN)	971	246	0.25
Research Aspect(RA)	635	159	0.25
Technology(Tech)	357	87	0.24
Method(Me)	221	55	0.25
Property(Prop)	358	89	0.25
Property Value(PV)	234	60	0.25
Experiment Name(EN)	225	58	0.25
Experiment Condition(EC)	373	89	0.24
Condition Value(CV)	418	111	0.26
Experiment Output(EO)	177	46	0.26
Equipment Used(EU)	221	54	0.24
Involved Element(IE)	183	49	0.26
Applicable Scenario(AS)	95	24	0.25
Sum	**4468**	**1127**	**0.25**

Table 2
Precision, Recall, F1-score, Average Value (AVG), and Standard Deviation (SD) of different models on the test set. In the contrast experiment, the downstream network of all language models is BiLSTM-CRF. The static language models used include Word2Vec, Mat2Vec and Fasttext, and the dynamic language models used include BERT, SciBERT, ALBERT, ClinicalBERT, BioBERT, MatBERT, MatSciBERT and MatTPUSciBERT. According to the proposed fusion strategy, four static and dynamic fusion models are obtained, namely BERT+Word2Vec, BERT+Fasttext, SciBERT+Word2Vec and SciBERT+Fasttext. Our model (SciBERT+Fasttext) F1 score is only 0.14 lower than the best model (MatSciBERT). The advantage of our model is that it does not need to use a large amount of material science literature to fine-tune or train the language model.

Model type	Language model	P_{ner}	R_{ner}	$F1_{ner}$	AVG	SD
Static	Word2Vec(2017)	68.66	66.08	67.35	67.36	1.29
	Mat2Vec(2019)	69.68	69.76	69.73	69.72	0.04
	Fasttext(2020)	69.04	67.56	68.29	68.30	0.75
Dynamic	BERT(2018)	78.25	77.01	77.62	77.63	0.62
	SciBERT(2019)	78.67	79.28	78.98	78.98	0.31
	ALBERT(2019)	77.96	75.83	76.88	76.89	1.06
	ClinicalBERT(2019)	77.15	75.96	76.55	76.55	0.60
	BioBERT(2020)	77.53	77.24	77.38	77.38	0.15
	MatBERT(2021)	78.83	80.11	79.46	79.47	0.64
	MatSciBERT(2022)	79.96	80.48	**80.22**	**80.22**	0.26
	MatTPUSciBERT(2022)	78.98	80.42	79.69	79.70	0.72
Our	BERT+Word2Vec	79.29	79.34	79.32	79.32	**0.03**
	BERT+Fasttext	76.12	79.80	77.91	77.94	1.84
	SciBERT+Word2Vec	78.09	**82.02**	80.01	80.04	1.96
	SciBERT+Fasttext	**80.35**	79.80	**80.08**	80.08	0.27

property (PRO), symmetry/phase label (SPL) and sample descriptor (DSC). The definition of these entities is inspired by the famous material science tetrahedron ‘processing’, ‘structure’, ‘characteristics’ and ‘property’. The annotation strategy of MatData is also BIO structure. Table 4 compares the F1 scores of six methods on individual entities (MAT, SMT, CMT, APL, PRO, SPL and DSC) and the overall F1 score (Sum). The experimental results show that our method achieves an F1 score of 88.16 on MatData dataset, while the Weston’s method [9] achieves an F1 score of 87.04. Meanwhile, we conducted experiments on the MatData using Mat2Vec, MatBERT, MatSciBERT and MatTPUSciBERT, which combined with BiLSTM-CRF to form the NER method. Mat2Vec achieved an F1 score of 80.19, MatBERT achieved an F1 score of 87.56, MatSciBERT achieved an F1 score of 88.39, and MatTPUSciBERT achieved an F1 score of 87.98. Compared to other methods in the experiment, our approach achieves near-optimal performance without the need for extensive language model training on large corpora.

4.2. Material composition table extraction

This work proposes a table recognition and composition extraction method that can obtain material compositions from tables in scientific literature in a structured form. 1327 material composition tables are collected from stainless steel literature for experiments. In order to improve the machine readability of the synthesis table extraction results, the extraction results are not directly presented in text form, but stored in json files using key–value structures. Taking a composition table as an example, the extraction result of our method is as follows: {material name: aisi 316 stainless steel, unit: wt%, composition: {(Si,0.22), (Mn, 2.63), (Cr,17.7), (C, 0.05), (Ni, 14.9), (S, 0.02), (Mo, 1.11)}}. We evaluate the extraction performance of the composition table using Algorithm 1. Finally, the information similarity score of the table recognition and composition extraction method proposed in this paper measured on 1327 material composition tables is 93.59%. An analysis is conducted on erroneous or inconclusive outcomes, revealing that the primary cause of erroneous results is attributed to issues with text recognition, whereas inconclusive outcomes predominantly occurred in composition tables with multiple rows of headers.

4.3. Property trend prediction

We conduct four groups of independent experiments on four datasets which are detailed in Section 3.3, using Decision Tree (DT), Random Forest (RF), K-Nearest Neighbor (KNN), AdaBoost and GBDT algorithms to train the property prediction models corresponding to the datasets. Based on each dataset, we use ten-fold cross-validation to calculate the evaluation metrics Precision (P_{prop}), Recall (R_{prop}) and F1-Score ($F1_{prop}$) for DT, RF, KNN, AdaBoost and GBDT respectively. The calculation formula is shown in Eqs. (7)–(9) and the experimental results are shown in Table 5.

$$P_{prop} = \frac{TP}{TP + FP} \tag{7}$$

$$R_{prop} = \frac{TP}{TP + FN} \tag{8}$$

$$F1_{prop} = \frac{2P_{prop}R_{prop}}{P_{prop} + R_{prop}} \tag{9}$$

where TP indicates correctly predicting positive samples as positive, FN indicates incorrectly predicting positive samples as negative, FP indicates incorrectly predicting negative samples as positive, and TN indicates correctly predicting negative samples as negative. Among these algorithms, GBDT has ideal applicability to low-dimensional data and strong robustness to outliers. Through linear combination of basis functions, multiple decision trees are integrated. During training, the residuals are continuously reduced. Finally, the final decision is formed based on the decision of each decision tree. Therefore, we use GBDT to train four property prediction models on four datasets and all models are initialized with GBDT default parameters. The parameter information is shown in Table 6. The corrosion resistance prediction model $F1_{prop}$ is 81.02, the ductility prediction model $F1_{prop}$ is 83.51, the strength prediction model $F1_{prop}$ is 80.13, and the hardness prediction model $F1_{prop}$ is 80.23. Each model can predict the corresponding property change trend according to the input material composition and technology.

In addition, to verify the effectiveness of the combination of text and table information, our work uses text information or table information alone for property prediction. The results are shown in Tables 7 and 8. A comparison of scores for property prediction using the GBDT on text and table data, only text data and only table data is shown in Fig. 8. The combination of text and table information applied to the prediction of properties is significantly more effective than only text or table information.

Table 3
F1-score of thirteen entity label categories for different models on the test set. In the contrast experiment, the downstream network of all language models is BiLSTM-CRF. The table header shows 13 entity categories, and the first column lists different language models. 'W2V' stands for Word2Vec model; 'M2V' stands for Mat2Vec; 'Fast' stands for Fasttext model; 'BE' stands for BERT model which is base version (L=12, H=768, A=12, Total Parameters=110M); 'Sci' stands for SciBERT; 'AL' stands for ALBERT; 'Clin' stands for ClinicalBERT; 'Bio' stands for BioBERT; 'MB' stands for MatBERT; 'MSB' stands for MatSciBERT; 'MTSB' stands for MatTPUSciBERT. Our models (BE+W2V, BE+Fast, Sci+W2V and Sci+Fast) use the dynamic and static word vector fusion strategy to obtain the highest F1-Score among the seven entity categories.

Model	MN	RA	Tech	Me	Prop	PV	EN	EC	CV	EO	EU	IE	AS
W2V	71.97	53.11	70.15	53.33	73.11	67.58	81.54	48.28	81.61	71.43	71.61	65.95	60.47
M2V	75.00	53.58	71.75	61.44	77.35	71.72	86.61	55.22	78.65	71.70	65.38	67.41	70.83
Fast	71.41	54.20	65.22	60.99	74.66	71.94	85.48	54.29	80.83	72.00	64.20	70.42	59.52
BE	83.81	65.64	77.99	**76.33**	74.18	75.94	86.17	55.95	85.58	75.58	74.09	77.46	68.18
Sci	85.17	61.92	85.19	70.45	80.23	70.88	88.79	58.62	86.46	79.19	76.54	77.03	70.41
AL	82.15	62.48	80.75	72.13	78.21	67.84	87.41	59.31	84.55	70.99	69.10	80.53	64.46
Clin	82.17	66.51	82.38	67.16	79.42	69.52	88.42	57.36	82.18	71.37	66.86	81.42	57.58
Bio	81.82	66.58	78.96	66.76	79.62	69.19	88.32	59.36	85.99	74.34	71.72	81.48	68.57
MB	86.05	66.13	**84.16**	70.14	78.81	72.91	88.54	58.13	85.09	79.10	76.00	83.33	72.27
MSB	85.16	66.54	83.15	70.99	**82.12**	71.29	88.89	**63.41**	87.26	81.95	76.81	82.19	69.11
MTSB	86.85	61.92	84.19	71.72	79.49	72.54	89.07	60.71	87.87	76.61	**78.15**	75.90	**69.70**
BE+W2V	85.92	63.56	79.82	68.06	78.42	73.80	90.42	57.58	89.41	83.19	80.99	78.72	66.67
BE+Fast	82.83	**66.78**	82.77	69.56	78.85	73.19	**91.81**	59.09	86.34	72.03	74.89	77.87	62.99
Sci+W2V	85.80	65.69	82.10	72.50	78.87	**83.16**	90.99	61.17	86.06	82.41	77.53	78.32	73.44
Sci+Fast	**87.80**	63.34	79.47	70.59	78.63	71.68	88.97	56.91	**89.31**	**82.74**	74.37	**82.65**	62.30

Table 4
On the material NER data set MatData, F1-score extracted by six methods for seven entities. MatData materials NER dataset defines seven categories of materials entities: inorganic material (MAT), synthesis method (SMT), characterization method (CMT), material application (APL), material property (PRO), symmetry/phase label (SPL) and sample descriptor (DSC).

Model	MAT	SMT	CMT	APL	PRO	SPL	DSC	Sum
Mat2Vec	87.65	68.85	79.08	74.89	74.47	61.38	78.88	80.19
MatBERT	92.70	79.58	87.79	79.58	81.45	84.21	88.83	87.56
MatSciBERT	92.03	83.24	**88.95**	84.33	81.06	81.11	89.71	88.39
MatTPUSciBERT	92.20	**83.49**	86.23	**86.22**	80.63	82.52	88.65	87.98
Weston [9]	90.30	81.37	86.01	80.63	**83.19**	82.05	**92.13**	87.04
Our	**92.71**	81.89	87.81	79.96	82.76	**84.57**	91.32	88.16

Table 5
Precision, Recall and F1-Score of different machine learning algorithms in predicting trends in corrosion resistance, ductility, strength and hardness. Data used are from texts and tables in the scientific literature.

Algorithm	Corrosion resistance			Ductility			Strength			Hardness		
	P_{prop}	R_{prop}	$F1_{prop}$	P_{prop}	R_{prop}	$F1_{prop}$	P_{prop}	R_{prop}	$F1_{prop}$	P_{prop}	R_{prop}	$F1_{prop}$
DT	76.29	60.42	65.49	71.44	67.18	60.62	55.05	**91.42**	66.31	79.61	52.41	62.20
RF	57.27	82.46	58.57	60.33	**90.20**	71.66	57.04	83.51	64.03	69.25	66.97	67.13
KNN	72.85	78.16	74.72	67.79	78.14	71.91	74.43	83.51	78.60	70.28	75.82	72.17
AdaBoost	**80.13**	**84.41**	**81.93**	79.86	84.34	81.66	**79.26**	81.70	80.10	74.77	74.91	74.52
GBDT	79.93	84.38	81.02	**82.45**	85.06	**83.51**	75.07	86.20	**80.13**	**82.92**	**78.59**	**80.23**

Table 6
Details of GBDT model.

Parameter	Value	Parameter	Value	Parameter	Value
Loss	deviance	learning_rate	0.1	n_estimators	100
Criterion	friedman_mse	min_samples_split	2	min_samples_leaf	1
max_depth	3	min_weight_fraction_leaf	0.0	verbose	0
validation_fraction	0.1	min_impurity_decrease	0.0	tol	0.001

Table 7
Precision, Recall and F1-Score of different machine learning algorithms in predicting trends in corrosion resistance, ductility, strength and hardness. Data used are from texts in the scientific literature. Compared with Table 5, '↑' means higher score and '↓' means lower score.

Algorithm	Corrosion resistance			Ductility			Strength			Hardness		
	P_{prop}	R_{prop}	$F1_{prop}$	P_{prop}	R_{prop}	$F1_{prop}$	P_{prop}	R_{prop}	$F1_{prop}$	P_{prop}	R_{prop}	$F1_{prop}$
DT	84.29↑	48.98↓	61.43↓	65.34↓	66.79↓	51.86↓	36.44↓	70.00↓	47.89↓	65.16↓	47.99↓	50.42↓
RF	62.37↑	80.87↓	67.83↑	57.04↓	91.52↑	69.69↓	39.33↓	80.00↓	52.49↓	64.49↓	65.66↓	64.46↓
KNN	58.73↓	74.53↓	64.35↓	66.18↓	67.69↓	63.94↓	61.55↓	72.68↓	65.52↓	66.47↓	68.97↓	65.31↓
AdaBoost	58.97↓	87.35↑	69.80↓	71.73↓	67.66↓	68.98↓	60.66↓	79.45↓	68.06↓	66.97↓	50.05↓	54.79↓
GBDT	66.33↓	85.63↑	74.19↓	76.09↓	70.71↓	72.96↓	61.88↓	82.46↓	70.28↓	77.23↓	65.16↓	69.16↓

5. Applications

In this section, we apply material science literature text and table information extraction methods to 11058 stainless steel scientific papers. We mine 13 types of material entities from the scientific literature text, resulting in 2.36 million material entities. We use the text mining results to analyze the research hotspots, trend changes and potential relationships between some entity categories of stainless steel

Table 8
Precision, Recall and F1-Score score of different machine learning algorithms in predicting trends in corrosion resistance, ductility, strength and hardness. Data used are from tables in the scientific literature. Compared with Table 5, '↑' means higher score and '↓' means lower score.

Algorithm	Corrosion resistance			Ductility			Strength			Hardness		
	P_{prop}	R_{prop}	$F1_{prop}$	P_{prop}	R_{prop}	$F1_{prop}$	P_{prop}	R_{prop}	$F1_{prop}$	P_{prop}	R_{prop}	$F1_{prop}$
DT	78.80↑	52.37↓	61.01↓	75.82↑	61.33↓	59.99↓	56.22↑	82.40↓	58.72↓	68.29↓	59.25↑	58.94↓
RF	70.05↑	69.85↓	67.63↑	63.53↑	83.94↓	71.34↑	58.58↑	81.61↓	60.79↓	62.79↓	71.57↑	64.98↓
KNN	73.21↑	69.61↓	70.21↓	70.30↑	75.64↓	71.97↑	69.02↓	74.51↓	70.72↓	68.60↓	60.54↓	63.62↓
AdaBoost	77.81↓	75.54↓	76.27↓	76.40↓	86.42↑	80.46↓	68.12↓	73.92↓	70.30↓	68.79↓	70.04↓	68.94↓
GBDT	80.30↑	79.27↓	79.57↓	83.78↑	83.74↓	82.31↓	75.15↑	76.47↓	75.55↓	75.14↓	77.00↓	75.15↓

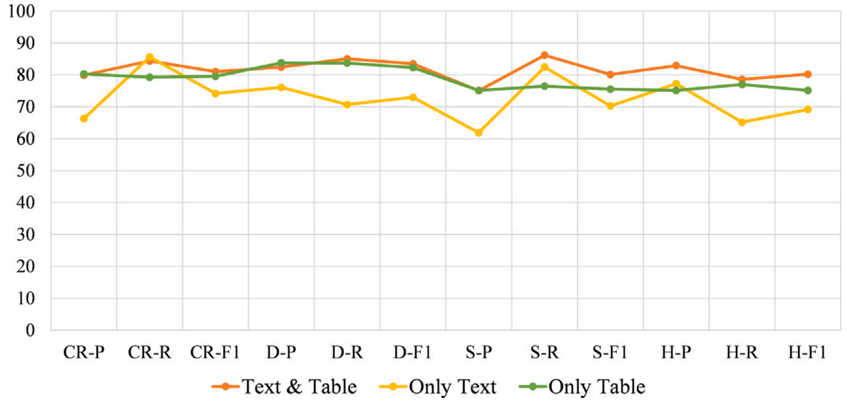


Fig. 8. Score comparison of property predictions using the GBDT on text & table data, only text data and only table data. 'CR' stands for corrosion resistance; 'D' stands for ductility; 'S' stands for strength; 'H' stands for hardness. 'P' stands for precision; 'R' stands for recall; 'F1' stands for F1 score. 'CR-P' means the precision of the model in predicting corrosion resistance, and so on.

from 2012 to 2021. In addition, we extract composition tables from 11058 papers and obtain 7970 sets of material composition information. Finally, material composition, technology, property and property trend data are selected from 2.36 million material entities and 7970 material composition information are obtained. Based on these data, four property prediction models are trained using GBDT algorithm. These models can predict the property change trends of stainless steel corrosion resistance, ductility, strength and hardness according to the composition, technology and property category.

5.1. Single entity statistics

In this subsection, we analyze the trend of the research heat of the 11 entities of material name, research aspect, technology, method, property, experiment name, experimental condition, experiment output, involved element, equipment used and applicable scenario between 2012 and 2021. The distributions of temperature, time, scan rate, voltage, laser power, heating rate, cooling rate and load condition values in different years are collected and displayed using scatter plot. In addition, the distributions of 'Temperature-Time-Year' and 'Load-Time-Year' are also counted. For the property and property value, we plot the distribution of the values of tensile strength and yield strength over different years.

Material Name & Year: This work calculates the proportions of the occurrence frequencies of austenitic stainless steel, martensitic stainless steel, ferritic stainless steel, duplex stainless steel and precipitation hardening stainless steel in scientific literature from 2012 to 2021. The specific types of stainless steel and the statistical results are shown in Fig. 9. The 316L stainless steel has received more attention from researchers in recent years, with the frequency ratio increasing from 0.33 in 2012 to 0.38 in 2021. 316L is a low-carbon stainless steel with added Mo element, which endows it with better corrosion resistance and high-temperature performance, especially suitable for seawater, brine and high-temperature environments. This is consistent with our statistical results of the application scenarios, which show that stainless steel is more widely used in nuclear power plants and aerospace fields. However, the research enthusiasm for 304 stainless steel is gradually decreasing, with the frequency ratio decreasing from 0.32 in 2012 to 0.21 in 2021. In addition, the attention of 316 stainless steel, carbon steel, super duplex stainless steel, etc., is increasing. The statistical result graph displays the changes of attention degree of different types of stainless steel in the past 10 years, revealing the research trends and hotspots in the field of stainless steel. It can provide reference for relevant researchers to grasp the future research and development direction.

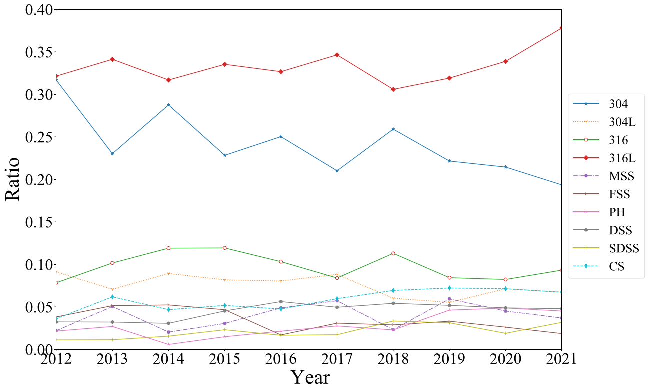


Fig. 9. Trends in materials. '304' is 'Austenite 304 SS', '304L' is 'Austenite 304L SS', '316' is 'Austenite 316 SS', '316L' is 'Austenite 316L SS', 'MSS' is 'Martensite (410&420&440) SS', 'FSS' is 'Ferrite (409&430&446) SS', 'PH' is '17–4 PH SS', 'DSS' is 'DSS (UNS S31500&S31803)', 'SDSS' is 'SDSS (UNS S32750&S32550)' and 'CS' is 'Carbon Steel'.

Research Aspect & Year: This work conducts data on 10 research aspects such as microstructure, corrosion, coating, welding, etc. and detailed statistical data shows in Fig. 10. The corrosion has been a research hotspot, and the frequency ratio is maintained between 0.25–0.30. Stainless steel has a passive film on its surface that prevents the metal from reacting with the environment, thereby maintaining its

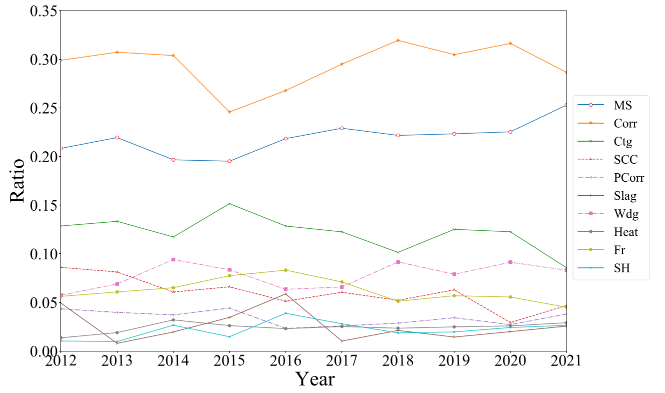


Fig. 10. Trends in research aspects. ‘MS’ is ‘Microstructure’, ‘Corr’ is ‘Corrosion’, ‘Ctg’ is ‘Coating’, ‘SCC’ is ‘Stress Corrosion Cracking’, ‘PCorr’ is ‘Pitting Corrosion’, ‘Slag’ is ‘Slag’, ‘Wdg’ is ‘Welding’, ‘Heat’ is ‘Heat Input/Transfer’, ‘Fr’ is ‘Friction’ and ‘SH’ is ‘Strain Hardening’.

properties and appearance. Therefore, its corrosion resistance has been a hot topic of research. In recent years, the study of microstructure has received increasing attention, and the study of surface coating has shown a downward trend. The microstructure of stainless steel has a significant impact on its properties. Studying the microstructure can provide important information for the improvement of stainless steel materials. Therefore, more and more scientists in the field of materials science have started to study the microstructure. Statistical analysis of the popularity of different research aspects can reveal the current hot topics in stainless steel materials. Researchers can select their subsequent research directions based on the statistical information, thereby improving the research efficiency.

Technology & Year: This work collects data for 7 technologies such as anneal, heat treatment, selective laser melting, laser powder bed fusion, etc. and Fig. 11 has more information. The technology of anneal is continuously decreasing, especially after 2018, and the frequency ratio declines continuously from 0.4 to 0.2. Annealing treatment makes the structure of stainless steel tend to equilibrium state, reducing its corrosion resistance. At the same time, annealing treatment also makes the hardness of stainless steel too low, which is not suitable for high strength and high wear resistance occasions. Therefore, the frequency of annealing process in stainless steel scientific literature has decreased significantly. Laser powder bed fusion technology is closely related to additive manufacturing. Additive manufacturing technology is an emerging technology of rapid manufacturing, which attracts the attention of researchers. Therefore, laser powder bed fusion technology is also widely mentioned, and the ratio has reached 0.2 in 2021.

Property & Year: For stainless steel, this paper collects 8 kinds of property such as hardness, corrosion resistance, yield strength, tensile strength, etc. The statistical result is shown in Fig. 12. The corrosion resistance of stainless steel has always been an important topic, but its research interest has declined in recent years. In contrast, more and more materials science literature involves the strength of stainless steel. The reason for this change is that the application requirements have changed. The strength of stainless steel is closely related to its wear resistance, durability and stability. In industrial manufacturing (such as automobile and aircraft manufacturing), lightweight and high-strength stainless steel materials are used to improve the load-bearing capacity and fatigue resistance of the body and fuselage. Our statistical results can make researchers pay attention to this change in property research.

Experiment Name & Year: In the scientific literature, the experimental part contains a lot of important information, among which the experiment name is the most intuitive description for experiment. The number of occurrences of tensile test, fatigue test and other 5 experiment names are counted. Detailed data are shown in Fig. 13. From the figure, we can find that the percentage of tensile test has been the largest which shows a continuous increasing trend, and the frequency percentage has reached 0.4 in 2021. A tensile test is a method of determining material characteristics under axial tensile loading, which provides data on strength and ductility of materials under uniaxial tensile forces. Tensile test also helps ensure that welding meets the required levels of strength and toughness, and provides critical information for selecting the best filler material. Therefore, tensile test is widely used in the study of mechanical properties of metal materials.

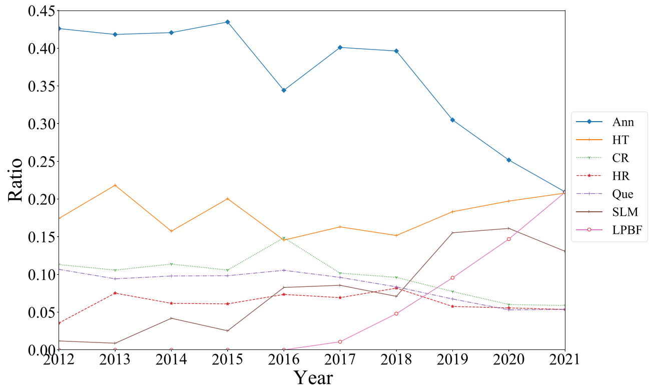


Fig. 11. Trends in technologies. ‘Ann’ is ‘Anneal’, ‘HT’ is ‘Heat Treatment’, ‘CR’ is ‘Cold Roll’, ‘HR’ is ‘Hot Roll’, ‘Que’ is ‘Quench’, ‘SLM’ is ‘Selective Laser Melting’ and ‘LPBF’ is ‘Laser Powder Bed Fusion’.

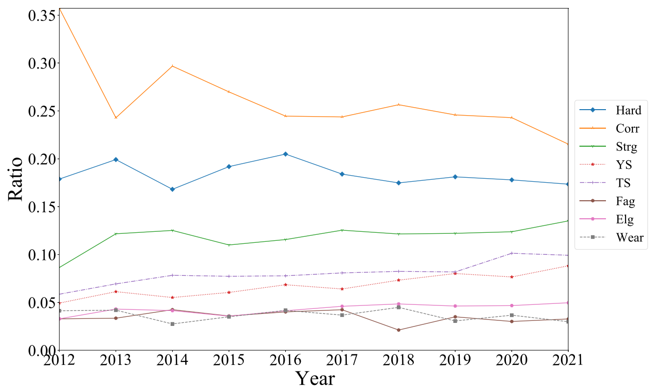


Fig. 12. Trends in properties. ‘Hard’ is ‘Hardness’, ‘Corr’ is ‘Corrosion Resistance’, ‘Strg’ is ‘Strength’, ‘YS’ is ‘Yield Strength’, ‘TS’ is ‘Tensile Strength’, ‘Fag’ is ‘Fatigue Resistance’, ‘Elg’ is ‘Elongation’ and ‘Wear’ is ‘Wear Resistance’.

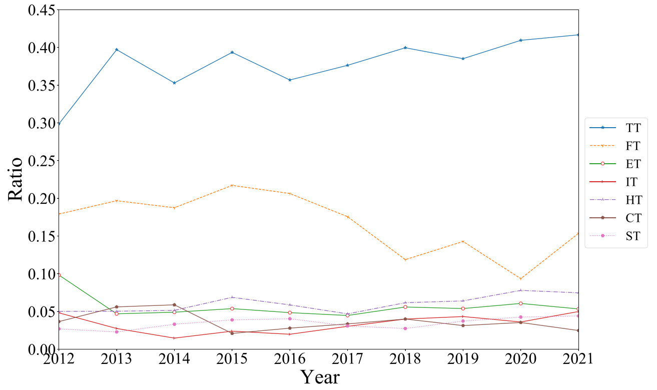


Fig. 13. Trends in experiment names. ‘TT’ is ‘Tensile Test’, ‘FT’ is ‘Fatigue Test’, ‘ET’ is ‘Electrochemical Test’, ‘IT’ is ‘Immersion Test’, ‘HT’ is ‘Hardness Test’, ‘CT’ is ‘Creep Test’ and ‘ST’ is ‘Shear Test’.

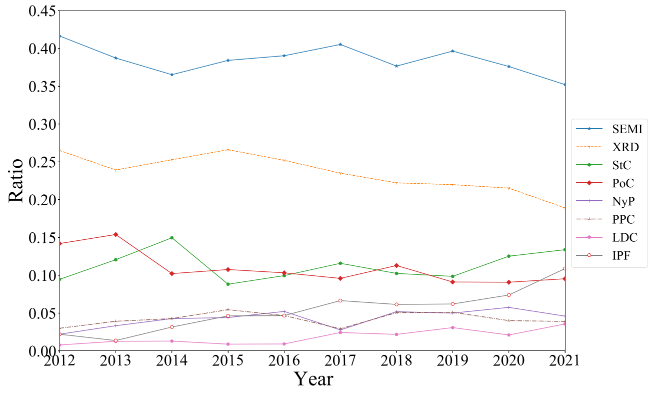


Fig. 14. Trends in experiment outputs. ‘SEMI’ is ‘Scanning Electron Microscopy Image’, ‘XRD’ is ‘X-ray Diffraction Pattern’, ‘StC’ is ‘Stress–strain Curve’, ‘PoC’ is ‘Polarization Curve’, ‘NyP’ is ‘Nyquist Plot’, ‘PPC’ is ‘Potentiodynamic Polarization Curve’, ‘LDC’ is ‘Load-Displacement Curve’ and ‘IPF’ is ‘Inverse Pole Figure’.

Experiment Output & Year: The experimental results are usually presented in the form of graphs. Various graphical information from experiments is collected, which includes scanning electron microscopy image, X-ray diffraction pattern, etc., and we present the result in Fig. 14. The image of scanning electron microscopy is obtained by Scanning Electron Microscopy (SEM) and has been the most important form of experimental results, and the frequency ratio has been maintained above 0.35. SEM is an imaging technique that uses an electron beam to scan the surface of a sample, generating secondary electrons or backscattered electrons. Its resolution is higher than optical microscopy, reaching the nanometer level. SEM plays a very important role in material research, as it can be used to observe the morphology, interface condition, damage mechanism and material property prediction of various materials. In addition, more and more experimental results are presented by inverse pole figure that can be used to study the crystallographic texture of materials.

Equipment Used & Year: The number of occurrences of 7 devices was collected which includes scanning electron microscopy, optical microscopy, transmission electron microscopy, etc. The collected data is shown in Fig. 15. According to statistics, the experimental equipment used for studying stainless steel has remained consistent in the past decade. Scanning Electron Microscopy (SEM) is widely used by researchers, and the frequency ratio has been maintained above 0.45. SEM, Optical Microscope (OM) and Transmission Electron Microscope (TEM) are common observation techniques. SEM is suitable for observing surface morphology, OM is suitable for observing color and structure under low magnification, and TEM is suitable for observing internal structure.

Applicable Scenario & Year: Statistical analysis of applicable scenario is performed, specifically for nuclear power plant, aerospace, chemical industries, biomedical applications, etc. As shown in Fig. 16, we can observe that the application of stainless steel in scientific literature shows large fluctuations, but nuclear power plants have always been the most important application scenario. In addition, stainless steel is increasingly used in aerospace and additive manufactured parts. Our statistical results can assist researchers in identifying research findings. For example, developing lightweight, high-strength stainless steel for aerospace applications, or using additive manufacturing methods to produce stainless steel parts.

Method & Year: This work counts the times of methods, including Electron Backscatter Diffraction (EBSD), X-ray Diffraction (XRD), Energy Dispersive Spectroscopy (EDS), etc., and finds that EBSD, XRD and EDS methods are most widely used. Especially EBSD, its proportion has been increasing year by year. The complete data is shown in Fig. 17. EBSD is a technique for quantitative microstructural analysis at millimeter to nanometer scales in Scanning Electron Microscopy (SEM). It can provide crystallographic information of materials, such as grain orientation, grain boundary angle, grain size, shape and distribution, texture, strain and phase identification. Moreover, EBSD can be combined with other techniques such as Energy Dispersive Spectrometer (EDS), Wavelength Dispersive Spectrometer (WDS), Cathodoluminescence (CL) etc., to achieve multimodal analysis and improve the efficiency and accuracy of material characterization. Because of these features, EBSD is increasingly used by researchers.

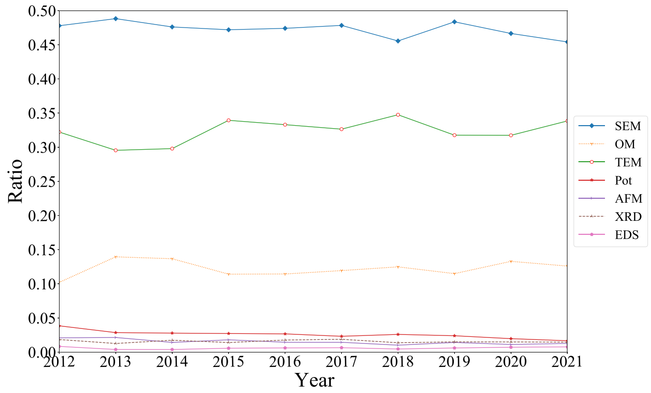


Fig. 15. Trends in equipments. ‘SEM’ is ‘Scanning Electron Microscopy’, ‘OM’ is ‘Optical Microscopy’, ‘TEM’ is ‘Transmission Electron Microscopy’, ‘Pot’ is ‘Potentiostat’, ‘AFM’ is ‘Atomic Force Microscopy’, ‘XRD’ is ‘X-ray Diffractometer’ and ‘EDS’ is ‘Energy Dispersive Spectrometer’.

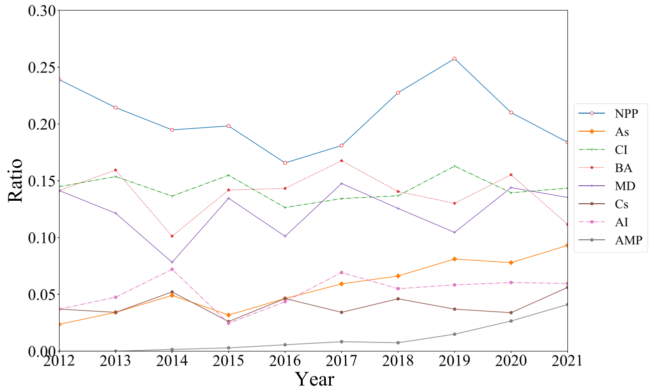


Fig. 16. Trends in applicable scenarios. ‘NPP’ is ‘Nuclear Power Plant’, ‘As’ is ‘Aerospace’, ‘CI’ is ‘Chemical Industries’, ‘BA’ is ‘Biomedical Applications’, ‘MD’ is ‘Medical Devices’, ‘Cs’ is ‘Construction’, ‘AI’ is ‘Automotive Industry’ and ‘AMP’ is ‘Additively Manufactured Parts’.

Experiment Condition & Year: The frequencies of 14 conditions are collected from 2012 to 2021. From Fig. 18, researchers can clearly understand the variation of the same condition in different years and the proportion of each condition in the same year. It is most intuitive to see that temperature has been the most important condition in the experiment. In metal material experiments, temperature condition affect process property (such as plasticity, hardening, etc.) and mechanical property (such as strength, hardness, etc.), and also cause special phenomena such as low-temperature brittleness and high-temperature creep. In addition, different temperature conditions lead to different thermal stress and thermal fatigue, as well as different oxidation and corrosion behaviors, which are directly related to the service life and feasibility of metals. Therefore, choosing appropriate temperature conditions in metal material experiments is very important, as it will directly affect the experimental results and analysis conclusions.

Involved Element & Year: The elements in the material composition have an important influence on the properties. This work counts 16 elements such as chromium (Cr), nitrogen (N), nickel (Ni), molybdenum (Mo), etc., and statistical result is shown in Fig. 19. Cr element

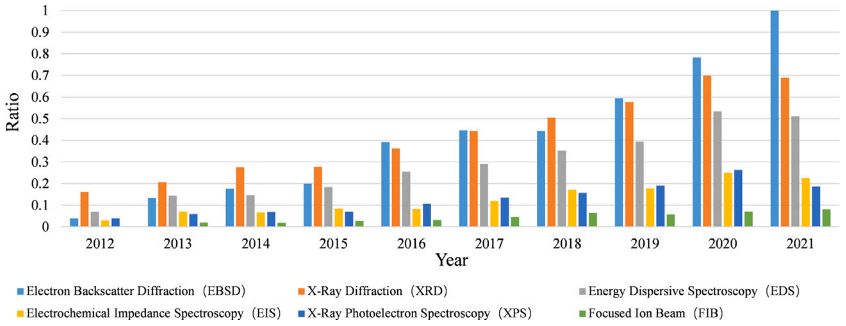


Fig. 17. Change of methods. The frequency of methods used in the experimental part of stainless steel scientific literature changed from 2012 to 2021.

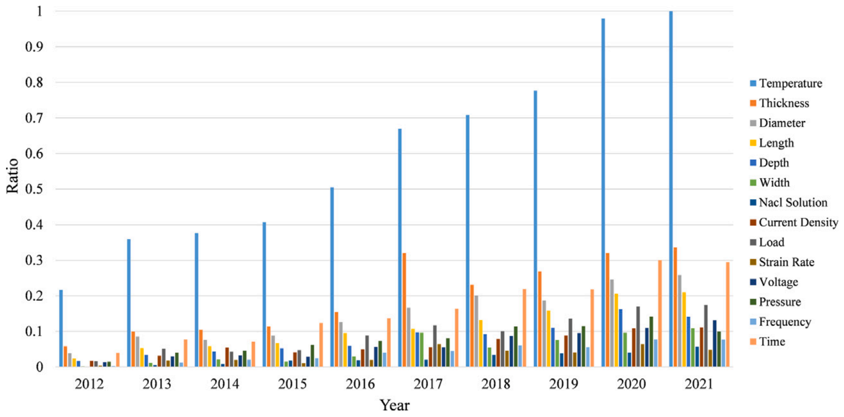


Fig. 18. Change of experiment conditions.

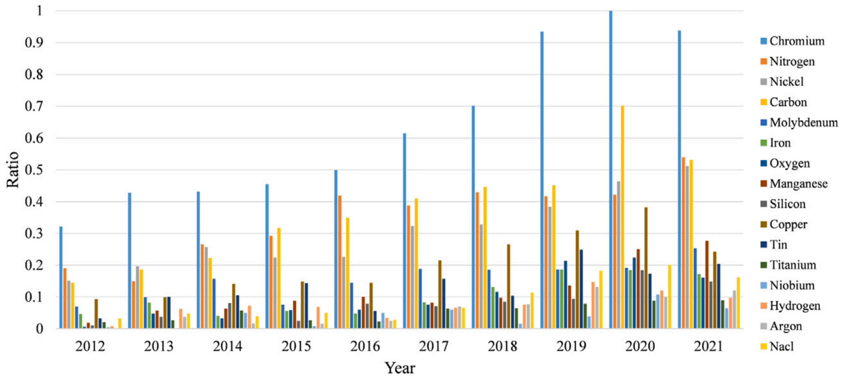


Fig. 19. Change of elements. The frequency of elements involved in stainless steel scientific literature from 2012 to 2021.

in stainless steel is the most concerned by researchers. Cr element is the main alloying element of stainless steel, and only when the Cr content reaches a certain value, the steel has corrosion resistance. The reason is that Cr element can cause a sudden change in the electrode potential of steel, from negative potential to positive electrode potential, thus inhibiting the oxidation reaction of iron in air. On the other hand, Cr element can form a dense layer of chromium oxide film (called passive film) on the surface of steel, which can prevent water and air from contacting with steel, thus protecting steel from further corrosion. Therefore, in stainless steel materials, Cr element is the key factor to improve corrosion resistance and protect metal luster.

Strength & Strength Value & Year: Tensile strength and yield strength are getting more and more attention from researchers, and we collect the values of these two strengths. 2100 tensile strength and yield strength values are collected and classified by year. The data is shown in Fig. 20 using the classified scatter chart. It can be clearly observed that the strength values mentioned in the literature are getting higher and higher. For example, the frequency of 1000 MPa for strength in 2021 is significantly increased compared with previous years.

5.2. Multi-label association analysis

From the extraction results of SFBC model, 674 (Property, Property Trend, Technology) ternary groups are collected during 2012–2021, and the ‘Property-Trend-Technology’ three-dimensional graph is drawn as Fig. 21(a) shows. The 1065 (Property, Property Trend, Element Involved) ternary groups are mined, and the ‘Property-Trend-Element’ three-dimensional graph is drawn as Fig. 21(b) shows. Figs. 21(a) and

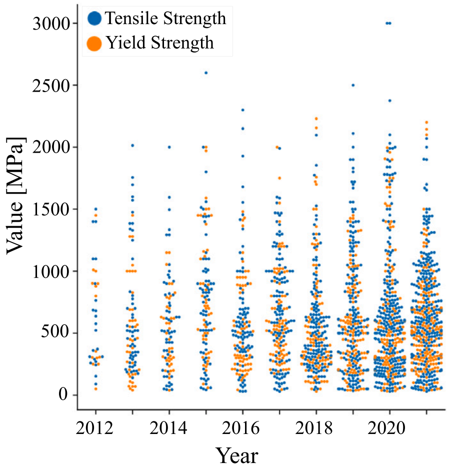


Fig. 20. Distribution of tensile strength and yield strength in stainless steel scientific literature from 2012 to 2021.

21(b) respectively illustrate the influences of seven manufacturing technologies on nine property changes and the effects of twelve elements on nine property changes based on statistical data. The color intensity indicates the frequency of performance improvement (shown in redder tones) or degradation (shown in bluer tones). The property attributes encompass corrosion resistance, hardness, wear resistance, oxidation resistance, tensile strength, toughness, ductility, yield strength, and biocompatibility. The property trends are described as improvement (↑) or degradation (↓). The manufacturing technologies include anneal, heat treatment, cold roll, hot roll, quench, selective laser melting, and laser powder bed fusion. The elements involved are chromium, nickel, molybdenum, nitrogen, carbon, copper, titanium, aluminum, iron, manganese, silicon, and cobalt. From Fig. 21(a), it can be observed that anneal and heat treatment technologies are closely correlated with improved corrosion resistance and ductility, as well as the deterioration of hardness, tensile strength, and yield strength. Fig. 21(b) reveals the significant role of chromium, nickel, and molybdenum in enhancing material corrosion resistance.

5.3. Property trend prediction

In addition to statistical analysis of literature mining results, we also construct four datasets of property prediction data from the mining results, and use GBDT algorithm to train a property prediction model on each set of data respectively. The four models trained can respectively predict the property trends of corrosion resistance, ductility, strength and hardness of stainless steel. The detailed content has been introduced in Sections 3.3 and 4.3.

6. Conclusion

This study proposes a method of information extraction and application for large-scale scientific literature in materials science, including text information extraction, composition table extraction and material property prediction. The aim of this paper is to assist in materials property improvement efforts, accelerate the pace of data-driven new material research, and promote the development of materials science. The SFBC model with a combination of GDWVs and DSWVs for texts in scientific literature is proposed. Using SFBC, a total of 2.36 million entities in 13 categories are extracted from 11,058 scientific papers, and the extraction results are counted from multiple perspectives to analyze the changes of research hotspots in stainless steel. The table recognition and composition extraction method is proposed for material composition tables. From composition tables, 7970 groups of material names, units, and compositions are extracted and stored in a structured form. From the material text and table extraction results, technologies, properties, property trends and material compositions are trained to obtain property change prediction models by GBDT. With prediction models, researchers can input material composition and technology to predict trend for property.

Compared with existing studies, our work does not limit itself to the extraction of textual content of scientific literature, but also covers the tabular data in the literature. On the basis of this, the extraction results of texts and tables are mined for deep association and eventually applied to material property prediction. This work is a further exploration of natural language processing and machine learning techniques in the field of materials science. However, materials science literature contains more than just texts and tables. Some non-text components such as images and formulae that provide key information also contain important information that can support and supplement the textual content. In our future work, we will do further research on multiple information association mining and extracting in science literature.

CRediT authorship contribution statement

Rui Zhang: Conceptualization, Methodology, Formal analysis, Writing – original draft, Writing – review & editing, Project administration. **Jiawang Zhang:** Conceptualization, Methodology, Software, Programming, Formal analysis, Writing – original draft, Writing – review & editing, Visualization, Preparation. **Qiaochuan Chen:** Conceptualization, Methodology, Formal analysis, Writing – original draft, Writing – review & editing. **Bing Wang:** Conceptualization, Methodology, Formal analysis, Writing – review & editing. **Yi Liu:** Validation, Data curation, Writing – review & editing. **Quan Qian:** Validation, Data curation, Writing – review & editing. **Deng Pan:** Validation, Data curation, Writing – review & editing. **Jinhua Xia:** Data curation, Writing – review & editing. **Yinggang Wang:** Data curation, Writing – review & editing. **Yuexing Han:** Conceptualization, Methodology, Formal analysis, Writing – original draft, Writing – review & editing, Project administration.

Declaration of competing interest

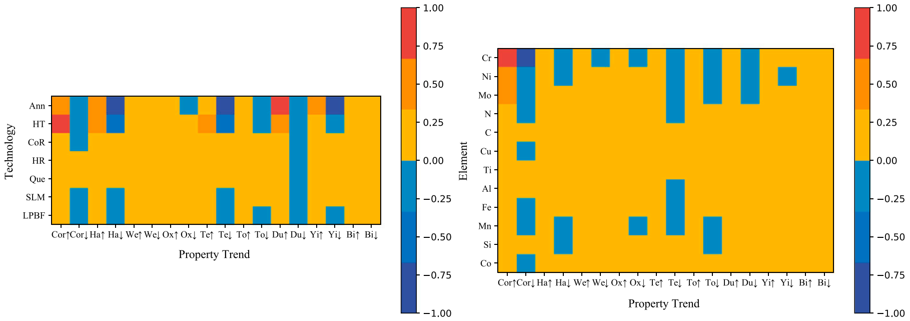
The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The statistical data can be available in Materials-Science-Literature-Mining at https://github.com/han-yuexing/Materials-Science-Literature-Mining. And our SFBC model code and stainless steel NER dataset have also been shared on Github. The Github project name is NER-SciBERT-Fasttext-BiLSTM-CRF, and researcher can visit project in https://github.com/han-yuexing/NER-SciBERT-Fasttext-BiLSTM-CRF.

Acknowledgments

This research is sponsored by the National Key Research and Development Program of China (Grant Nos. 2022YFB3707800, 2020YFB0704503), National Natural Science Foundation of China (Grant No. 52273228), Natural Science Foundation of Shanghai (Grant No. 20ZR1419000), Key Research Project of Zhejiang Laboratory (Grant No. 2021PE0AC02), Key Program of Science and Technology of Yunnan Province (Grant Nos. 202302AB080022, 202302AB080020, 202102AB080019-3), Key Project of Shanghai Zhangjiang National Independent Innovation Demonstration Zone (Grant No. ZJ2021-ZD-006), Shanghai Science and Technology Young Talents Sailing Program (Grant No. 23YF1412900).



(a) Property-Trend-Technology (b) Property-Trend-Element

Fig. 21. Thermodynamic Map of the Relationship between Technologies, Elements, and Property Trends. The color gradient in the legend indicates that the closer it is to red, the higher the frequency of performance improvement, while the closer it is to blue, the higher the frequency of performance decline. The X-axis of Figures (a) and (b) represents different property enhancements or declines, including Corrosion Resistance (Cor), Hardness (Ha), Wear Resistance (We), Oxidation Resistance (Ox), Tensile Strength (Te), Toughness (To), Ductility (Du), Yield Strength (Yi), and Biocompatibility (Bi). The property trends include improvement (↑) and decline (↓). The Y-axis in Figure (a) represents technologies, including Anneal (Ann), Heat Treatment (HT), Cold Roll (CoR), Hot Roll (HR), Quench (Que), Selective Laser Melting (SLM), and Laser Powder Bed Fusion (LPBF). The Y-axis in Figure (b) represents elements, including Cr, Ni, Mo, N, C, Cu, Ti, Al, Fe, Mn, Si, and Co.

References

[1] X.B. Qu, Research on the factors influencing the selection of new materials in product design, Design A (06) (2014) 11–12.

[2] J. Wei, X. Chu, X.Y. Sun, et al., Machine learning in materials science, InfoMat 1 (3) (2019) 338–358.

[3] K.T. Butler, D.W. Davies, H. Cartwright, et al., Machine learning for molecular and materials science, Nature 559 (7715) (2018) 547–555.

[4] W. Wu, Q. Sun, Applying machine learning to accelerate new materials development, Sci. Sin. Phys., Mech. Astron. 48 (10) (2018) 107001.

[5] M. Wang, T. Wang, P. Cai, et al., Nanomaterials discovery and design through machine learning, Small Methods 3 (5) (2019) 1900025.

[6] C. Nantasenamat, C. Isarankura-Na-Ayudhya, V. Prachayasittikul, Advances in computational methods to predict the biological activity of compounds, Expert Opin. Drug Discov. 5 (7) (2010) 633–654.

[7] S.P. Si, et al., Study on strengthening effects of Zr-Ti-Nb-O alloys via high throughput powder metallurgy and data-driven machine learning, Mater. Des. 206 (2021) 109777.

[8] B. Zhang, C.S. Yung, Data-driven phase recognition of steels for use in mechanical property prediction, Manuf. Lett. 30 (2021) 27–31.

[9] L. Weston, V. Tshitoyan, J. Dagdelen, et al., Named entity recognition and normalization applied to large-scale information extraction from the materials science literature, J. Chem. Inf. Model. 59 (9) (2019) 3692–3702.

[10] D. Westergaard, H.H. Staerfeldt, C. Tonsberg, et al., A comprehensive and quantitative comparison of text-mining in 15 million full-text articles versus their corresponding abstracts, PLoS Comput. Biol. 14 (2) (2018) e1005962.

[11] V. Venugopal, S. Sahoo, M. Zaki, et al., Looking through glass: Knowledge discovery from materials science literature using natural language processing, Patterns 2 (7) (2021) 100290.

[12] S. Guha, A. Mullick, J. Agrawal, et al., MatScIE: An automated tool for the generation of databases of methods and parameters used in the computational materials science literature, Comput. Mater. Sci. 192 (2021) 110325.

[13] F. Kuniyoshi, J. Ozawa, M. Miwa, Analyzing research trends in inorganic materials literature using NLP, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, Cham, 2021, pp. 319–334.

[14] S. Chandran, R. Kasturi, Structural recognition of tabulated data, in: Proceedings of 2nd International Conference on Document Analysis and Recognition, (ICDAR'93), IEEE, 1993, pp. 516–519.

[15] L. Hao, L. Gao, X. Yi, et al., A table detection method for pdf documents based on convolutional neural networks, in: 2016 12th IAPR Workshop on Document Analysis Systems, (DAS), IEEE, 2016, pp. 287–292.

[16] C. Tensmeyer, V.I. Morariu, B. Price, et al., Deep splitting and merging for table structure decomposition, in: 2019 International Conference on Document Analysis and Recognition, (ICDAR), IEEE, 2019, pp. 114–121.

[17] K. Rajan, Materials informatics: The materials gene and big data, Annu. Rev. Mater. Res. 45 (2015) 153–169.

[18] J.C. Mauro, A. Tandia, K.D. Vargheese, et al., Accelerating the design of functional glasses through modeling, Chem. Mater. 28 (12) (2016) 4267–4277.

[19] P. Bhaskar, R. Kumar, Y. Maurya, et al., Cooling rate effects on the structure of 45S5 bioglass: Insights from experiments and simulations, J. Non-Cryst. Solids 534 (2020) 119952.

[20] A. Ravi, Prediction of reduced glass transition temperature using machine learning, 2020, arXiv preprint arXiv:2005.08872.

[21] J. Xiong, T.Y. Zhang, S.Q. Shi, Machine learning of mechanical properties of steels, Sci. China Technol. Sci. 63 (7) (2020) 1247–1255.

[22] V.A. Hosseini, M. Thuvander, K. Lindgren, et al., Fe and Cr phase separation in super and hyper duplex stainless steel plates and welds after very short aging times, Mater. Des. 210 (2021) 110055.

[23] C. Zhao, Y. Bai, Y. Zhang, et al., Influence of scanning strategy and building direction on microstructure and corrosion behaviour of selective laser melted 316L stainless steel, Mater. Des. 209 (2021) 109999.

[24] T. Masumura, T. Tsuchiyama, Effect of carbon and nitrogen on work-hardening behavior in metastable austenitic stainless steel, Isij Int. 61 (2) (2021) 617–624.

[25] T.R. Tabrizi, M. Sabzi, S.H.M. Anijdan, et al., Comparing the effect of continuous and pulsed current in the GTAW process of AISI 316l stainless steel welded joint: Microstructural evolution, phase equilibrium, mechanical properties and fracture mode, J. Mater. Res. Technol. 15 (2021) 199–212.

[26] Q. Ma, C. Luo, S. Liu, et al., Investigation of arc stability, microstructure evolution and corrosion resistance in underwater wet fcaw of duplex stainless steel, J. Mater. Res. Technol. 15 (2021) 5482–5495.

[27] C. Zhang, J. Zhu, C. Ji, et al., Laser powder bed fusion of high-entropy alloy particle-reinforced stainless steel with enhanced strength, ductility, and corrosion resistance, Mater. Des. 209 (2021) 109950.

[28] S. Salahi, M. Kazemipour, A. Nasiri, Effects of microstructural evolution on the corrosion properties of AISI 420 martensitic stainless steel during cold rolling process, Mater. Chem. Phys. 258 (2021) 123916.

[29] J. Nie, L. Wei, Y. Jiang, et al., Corrosion mechanism of additively manufactured 316 L stainless steel in 3.5 wt% NaCl solution, Mater. Today Commun. 26 (2021) 101648.

[30] S.Y. Lee, C. Takushima, J. Hamada, et al., Macroscopic and microscopic characterizations of portevin-lechatelier effect in austenitic stainless steel using high-temperature digital image correlation analysis, Acta Mater. 205 (2021) 116560.

[31] T. Takai, T. Furukawa, H. Yamano, Thermophysical properties of austenitic stainless steel containing boron carbide in a solid state, Mech. Eng. J. 8 (4) (2021) 20-00540-20-00540.

[32] X. Zhang, S. Yang, J. Li, et al., Evolution of oxide inclusions in stainless steel containing yttrium during thermo-mechanical treatment, J. Mater. Res. Technol. 9 (3) (2020) 5982–5990.

[33] https://github.com/doccano/doccano.

[34] N. Reimers, I. Gurevych, Optimal hyperparameters for deep lstm-networks for sequence labeling tasks, 2017, arXiv preprint arXiv:1707.06799.

[35] https://github.com/pymupdf/PyMuPDF.

[36] I. Beltagy, K. Lo, A. Cohan, SciBERT: A pretrained language model for scientific text, 2019, arXiv preprint arXiv:1903.10676.

[37] E. Kim, Z. Jensen, A. van Grootel, et al., Inorganic materials synthesis planning with literature-trained neural networks, J. Chem. Inf. Model. 60 (3) (2020) 1194–1201.

[38] J. Devlin, M.W. Chang, K. Lee, et al., Bert: pre-training of deep bidirectional transformers for language understanding, 2018, arXiv preprint arXiv:1810.04805.

[39] T. Wolf, L. Debut, V. Sanh, et al., Transformers: state-of-the-art natural language processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2020, pp. 38–45.

[40] M. Schuster, K. Nakajima, Japanese and korean voice search, in: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP), IEEE, 2012, pp. 5149–5152.

[41] A. Joulin, E. Grave, P. Bojanowski, et al., Bag of tricks for efficient text classification, 2016, arXiv preprint arXiv:1607.01759.

[42] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.

[43] G.U.O. Zhi-xin, D. Xiao-long, Intelligent identification method of legal case entity based on BERT-BiLSTM-CRF, J. Beijing Univ. Posts Telecommun. 44 (4) (2021) 129.

[44] J. Redmon, A. Farhadi, Yolov3: An incremental improvement, 2018, arXiv preprint arXiv:1804.02767.

[45] B. Shi, X. Bai, C. Yao, An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition, IEEE Trans. Pattern Anal. Mach. Intell. 39 (11) (2016) 2298–2304.

[46] E. Kim, K. Huang, A. Tomala, et al., Machine-learned and codified synthesis parameters of oxide materials, Sci. data 4 (1) (2017) 1–9.

[47] https://github.com/PaddlePaddle/PaddleOCR.

[48] X. Zhong, E. ShafieiBavani, A. Jimeno Yepes, Image-based table recognition: data, model, and evaluation, in: European Conference on Computer Vision, Springer, Cham, 2020, pp. 564–580.

[49] J.H. Friedman, Greedy function approximation: A gradient boosting machine, Ann. Stat. (2001) 1189–1232.

[50] J.W. Xu, Y. Yang, A survey of ensemble learning approaches, J. Yunnan Univ. (Natural Sci. Edition) 40 (6) (2018) 1082–1092.

[51] P.I. Lixiang, C.U.I. Guimei, Optimizing GBDT's strip coiling temperature prediction with the evolutionary algorithm, J. South China Normal Univ.(Natural Sci. Edition) 54 (1) (2022) 122–127.

[52] J. Cheng, G. Li, X. Chen, Research on travel time prediction model of freeway based on gradient boosting decision tree, IEEE Access 7 (2018) 7466–7480.

[53] S. Deng, C. Wang, M. Wang, et al., A gradient boosting decision tree approach for insider trading identification: An empirical model evaluation of China stock market, Appl. Soft Comput. 83 (2019) 105652.

[54] L. Weston, V. Tshitoyan, J. Dagdelen, et al., Named entity recognition and normalization applied to large-scale information extraction from the materials science literature, J. Chem. Inf. Model. 59 (9) (2019) 3692–3702.

[55] Z. Lan, M. Chen, S. Goodman, et al., Albert: A lite bert for self-supervised learning of language representations, 2019, arXiv preprint arXiv:1909.11942.

[56] E. Alsentzer, J.R. Murphy, W. Boag, et al., Publicly available clinical BERT embeddings, 2019, arXiv preprint arXiv:1904.03323.

[57] J. Lee, W. Yoon, S. Kim, et al., BioBERT: A pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36 (4) (2020) 1234–1240.

[58] N. Walker, A. Trewartha, H. Huo, et al., The impact of domain-specific pre-training on named entity recognition tasks in materials science, 2021, Available at SSRN 3950755.

[59] T. Gupta, M. Zaki, N.M.A. Krishnan, MatSciBERT: A materials domain language model for text mining and information extraction, npj Comput. Mater. 8 (1) (2022) 102.

[60] https://huggingface.co/lfoppiano/MatTPUSciBERT.