


# Literature classification and its applications in condensed matter physics and materials science by natural language processing

Siyuan Wu, Tiannian Zhu, Sijia Tu, Ruijuan Xiao, Jie Yuan, Quansheng Wu, Hong Li, and Hongming Weng

**Citation:**Chin. Phys. B, 2024, 33 (**5**): 050704. DOI: 10.1088/1674-1056/ad3c30
Journal homepage: http://cpb.iphy.ac.cn; http://iopscience.iop.org/cpb

**What follows is a list of articles you may be interested in**

---

## Asymmetric magnetoimpedance effect and dipolar interactions of FINEMET/$SiO_2$/FePd composite ribbons

Yong-Bin Guo(郭永斌), Dao Wang(王岛), Zhong-Min Wang(王忠民), Lei Ma(马垒), and Zhen-Jie Zhao(赵振杰)
Chin. Phys. B, 2023, 32 (**7**): 070703. DOI: 10.1088/1674-1056/acad6e

## Photoreflectance system based on vacuum ultraviolet laser at 177.3 nm

Wei-Xia Luo(罗伟霞), Xue-Lu Liu(刘雪璐), Xiang-Dong Luo(罗向东), Feng Yang(杨峰), Shen-Jin Zhang(张申金), Qin-Jun Peng(彭钦军), Zu-Yan Xu(许祖彦), and Ping-Heng Tan(谭平恒)
Chin. Phys. B, 2022, 31 (**11**): 110701. DOI: 10.1088/1674-1056/ac89e4

## High-precision nuclear magnetic resonance probe suitable for *in situ* studies of high-temperature metallic melts

Ao Li(李傲), Wei Xu(许巍), Xiao Chen(陈霄), Bing-Nan Yao(姚冰楠), Jun-Tao Huo(霍军涛), Jun-Qiang Wang(王军强), and Run-Wei Li(李润伟)
Chin. Phys. B, 2022, 31 (**4**): 040706. DOI: 10.1088/1674-1056/ac4a70

## Design of a novel correlative reflection electron microscope for *in-situ* real-time chemical analysis

Tian-Long Li(李天龙), Zheng Wei(魏征), and Wei-Shi Wan(万唯实)
Chin. Phys. B, 2021, 30 (**12**): 120702. DOI: 10.1088/1674-1056/ac0522

## Efficient sampling for decision making in materials discovery

Yuan Tian(田原), Turab Lookman, and Dezhen Xue(薛德祯)
Chin. Phys. B, 2021, 30 (**5**): 050705. DOI: 10.1088/1674-1056/abf12d

---

**COMPUTATIONAL PROGRAMS FOR PHYSICS**

# Literature classification and its applications in condensed matter physics and materials science by natural language processing

Siyuan Wu(吴思远)[1,2,4], Tiannian Zhu(朱天念)[1,2,4], Sijia Tu(涂思佳)[1,3], Ruijuan Xiao(肖睿娟)[1,4,†], Jie Yuan(袁洁)[1,4], Quansheng Wu(吴泉生)[1,4], Hong Li(李泓)[1,‡], and Hongming Weng(翁红明)[1,4,§]

[1] *Institute of Physics, Chinese Academy of Sciences, Beijing* 100190, *China*
[2] *School of Physical Sciences, University of Chinese Academy of Sciences, Beijing* 100190, *China*
[3] *College of Materials Science and Optoelectronic Technology, University of Chinese Academy of Sciences, Beijing* 100049, *China*
[4] *Condensed Matter Physics Data Center of Chinese Academy of Sciences, Beijing* 100190, *China*

(Received 27 December 2023; revised manuscript received 27 March 2024; accepted manuscript online 9 April 2024)

The exponential growth of literature is constraining researchers' access to comprehensive information in related fields. While natural language processing (NLP) may offer an effective solution to literature classification, it remains hindered by the lack of labelled dataset. In this article, we introduce a novel method for generating literature classification models through semi-supervised learning, which can generate labelled dataset iteratively with limited human input. We apply this method to train NLP models for classifying literatures related to several research directions, i.e., battery, superconductor, topological material, and artificial intelligence (AI) in materials science. The trained NLP 'battery' model applied on a larger dataset different from the training and testing dataset can achieve F1 score of 0.738, which indicates the accuracy and reliability of this scheme. Furthermore, our approach demonstrates that even with insufficient data, the not-well-trained model in the first few cycles can identify the relationships among different research fields and facilitate the discovery and understanding of interdisciplinary directions.

**Keywords:** natural language processing, text mining, materials science

**PACS:** 07.05.Kf **DOI:** 10.1088/1674-1056/ad3c30

## 1. Introduction

In recent years, an exponential increase in published literatures has made it impossible to read and sort all interested works individually and timely. To address this challenge, an effective solution lies in leveraging natural language processing (NLP),[1,2] particularly large language models (LLMs) like BERT[3] and GPT[4] which have gained prominence. While pre-training has alleviated the need for high-quality text dataset, the scarcity of dataset remains a limitation for LLMs.[5] Thus, acquiring high-quality text entries at a low cost becomes imperative in advancing NLP for literature classification and generative AI the same as data in materials science.[6] As shown from ChatGPT, some bad data has also been understood by AI even the input has some wrong words. It means that data is important as well as the "wrong" data. What we need to pay attention to is how to get enough data.

Since Google put forward the concept of Word2Vec,[7] the word embeddings[8–11] have become the most popular method to train the relationship among words. After the word embeddings, there exist multi-layer perception (MLP), recurrent neural networks (RNN),[12] long short-term memory (LSTM)[13] and transformer[14] to capture the semantic meaning. MLP is the same as the logic gate in the integrated circuit which trains the weight to control the output. RNN[12] is designed to learn the sequential problem while it has the gradient vanishing problem and gradient exploding problem, so LSTM[13] is used to replace it. However, recurrent networks such as RNN and LSTM fail to capture the long text and parallel computing. Attention mechanism is presented[14] and pretrain becomes a necessary step for NLP such as BERT[3] and GPT.[4]

The tasks of NLP have been divided into two classifies: natural language understanding (NLU) and natural language generation (NLG). For the former task, the NLU needs to handle the text categorization, named entity recognition, question answering and similarity analysis. While the NLG needs to generate sentences including machine translation, summarization, dialogue, etc. BERT belongs to the former and GPT belongs to the latter. In recent six months (2023.8–2024.1), the LLMs have been applied in materials science.[15–28] One of the efforts is to use LLMs, especially GPT-3.5 and GPT-4, to generate datasets. For example, Yang Jeong Park *et al.*[15] generated 1.5 million materials narratives using ChatGPT. Samuel J. Yang *et al.*[16] generated extra DFT data using GPT-4 and Xiang Zhang *et al.*[17] did for phosphors data. Another way is to pretrain or fine-tune the LLMs for differ-

†Corresponding author. E-mail: rjxiao@iphy.ac.cn
‡Corresponding author. E-mail: hli@iphy.ac.cn
§Corresponding author. E-mail: hmweng@iphy.ac.cn

© 2024 Chinese Physical Society and IOP Publishing Ltd 

ent tasks. Andre Niyongabo Rubungo *et al.*[18] fined tune T5 to generate the crystal representation. Tatsuya Sagawa *et al.*[19] pre-trained T5 with the open reaction database as ReactionT5. Markus J. Buehler[20] fine-tuned MechGPT based on OpenOrca-Platypus2-13B to train in the mechanics of materials domain. Zhiling Zheng *et al.*[21–24] discovered novel MOFs by GPT-4. Daniil A. Boiko *et al.*[25] used GPT-4 to design chemical experiments. Xianjun Yang *et al.*[26] trained Quokka based on LLaMa-2. Ziyi Chen *et al.*[27] fine-tuned MatChat based on LLAMA-2 to predict inorganic material synthesis pathways. The ChatGPT in the material design has also been discussed in Ref. [28].

As one of the three crucial elements in machine learning, high-quality dataset continues to be deficient in scientific literature. Compounding the issue, these entries can only be labelled by scientific researchers. This creates a contradiction for NLP in scientific research: we require NLP models to streamline literature collecting and classifying efficiently, but the models necessitate sufficient labelled entries. Therefore, it becomes imperative to develop a rapid and effective method for data labelling and constructing NLP models tailored to the needs of scientific researchers.

In this article, we present an effective approach to generating 10k data entries with NLP models using semi-supervised learning based on TensorFlow.[29] Our estimation suggests that about 1000–4000 entries need to be labelled to create a suitable NLP model, involving approximately 10k entries for training and testing purposes. We proceed to build four models: 'battery', 'superconductor', 'topological material' and 'AI in materials science', which we validate with an additional dataset of 150k entries from Web of Science,[30] demonstrating impressive robustness of the trained models.

Furthermore, we emphasize the crucial role of the unreliable model at first few cycles during the development of NLP models, as it not only aids in creating effective models but also reveals intricate relationships among diverse research fields.

## 2. The concept and theory of text classification in NLP

### 2.1. Tokenize the sentence

The core of NLP lies in Word2Vec, which enables the transformation of human-readable sentences into computer-manageable vectors. Before Word2Vec, we need to transform the human-readable sentences into corresponding vectors which are also called as tokenizer.

Taking a sentence for example, 'I come from E01 group at the Institute of Physics, Chinese Academy of Sciences', we tokenize it into the following words: 'I', 'come', 'from', 'E01', 'group', 'at', 'the', 'Institute', 'of', 'Physics', 'Chinese', 'Academy', 'of', 'Sciences'. These words are then converted into the vector [0, 1, 2, 12, 3, 4, 5, 6, 7, 8, 9, 10, 7, 11], as illustrated in Fig. 1(a) using the dictionary in Fig. 1(b).

Similarly, for another two sentences, 'I come from T03 group at the Institute of Physics, Chinese Academy of Sciences' and 'I come from SC02 group at the Institute of Physics, Chinese Academy of Sciences', we can also transform them into vectors: [0, 1, 2, 13, 3, 4, 5, 6, 7, 8, 9, 10, 7, 11] and [0, 1, 2, 14, 3, 4, 5, 6, 7, 8, 9, 10, 7, 11], respectively.

Importantly, these three sentences can be differentiated by machine learning: they share the common part, 'Institute of Physics, Chinese Academy of Sciences', while the differing parts are the various groups labelled as E01, T03, and SC02.

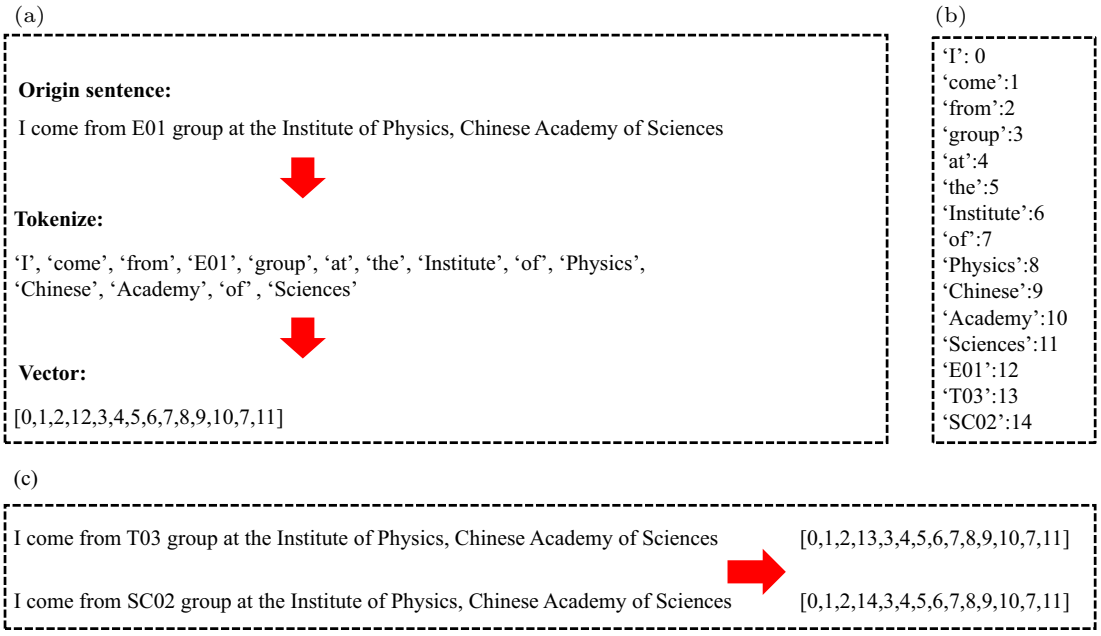



**Fig. 1.** (a) An example of transforming the natural sentence into a vector. (b) The dictionary for (a). (c) Another two examples for different sentences.

### 2.2. Word2Vec

The core of Word2Vec is to transform the word into a vector according for different contexts. What the Word2Vec or word embedding does is to classify different meanings for the same word. For example, 'Chinese' represents the Chinese people or the Chinese language. Using Word2Vec learning from the context we can distinguish it. In the above example, 'E' in 'E01' represents 'energy', 'SC' in 'SC02' represents 'superconductor', while 'T' in 'T03' represents 'theory'. The object of this step is to distinguish their meaning.

### 2.3. The neural network to classification

As mentioned before, we can use MLP, RNN, LSTM and even transformer to construct the neural network. Here in order to reduce the computation cost, we use only one-layer MLP to do the classification. All the parameters are default values in TensorFlow to reduce workload. For text classification, the input of the model is the text you need to handle while the output can be the Boolean value: true and false or 1 and 0 representing whether the text belongs to the category or not.

## 3. Constructing NLP models by semi-supervised learning

Here, we employ an emotion classification model based on TensorFlow (GPU v2.6. and CPU v2.10.0, the details are shown in Data available).[29] Despite the absence of explicit mentions of happiness or sadness in the text, humans tend to impart emotions to the sentences. In selecting the emotion classification model, we take a similar approach, recognizing that determining whether a text belongs to a specific field or not also involves human input. The results demonstrate the strong robustness and reliability of the models.

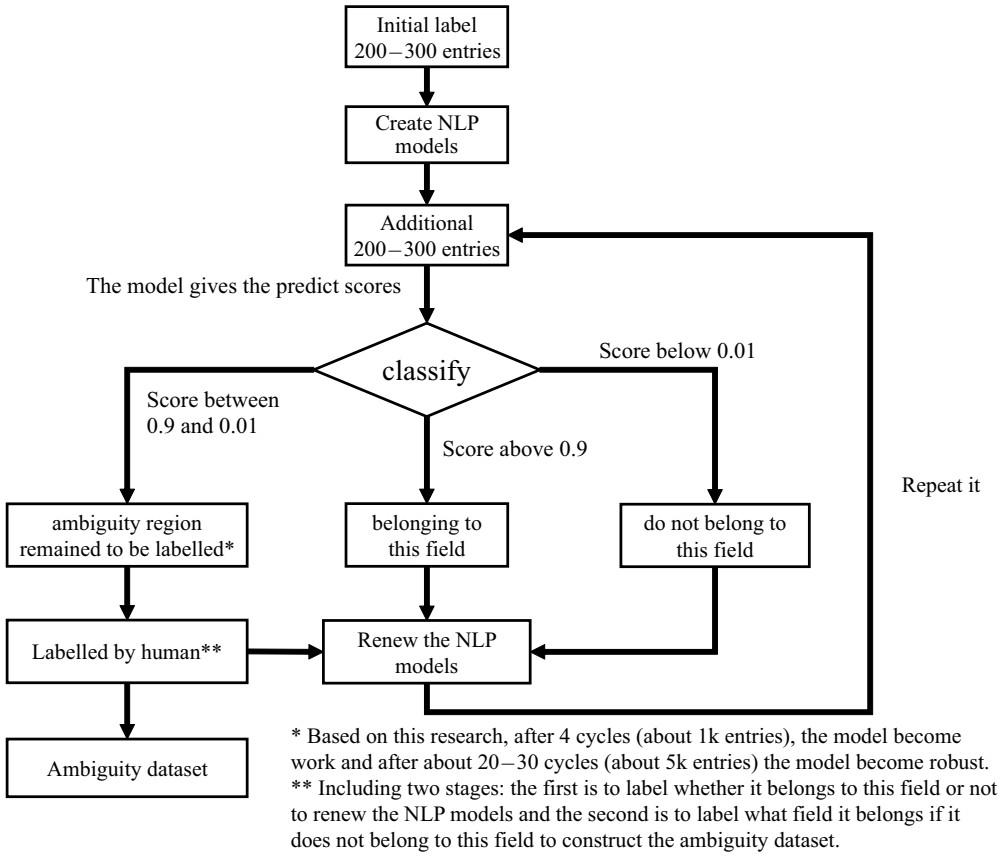



**Fig. 2.** The flow chart for labelling the entries and creating the NLP model for a target field.

The dataset is constructed following the representation in Fig. 2. Initially, we select about 200–300 entries labelled by human and build the model based on the emotion classification model. This newly built model may exhibit bias in classifying literatures, but this is not a major concern. Using the biased model, we predict an additional dataset consisting of 200–300 entries. Among these entries, approximately 5% of them score above 0.01 (here, '1' represents the literature that belongs to the target field, while '0' does not; '0.1' represents a 10% probability of belonging to the target field). Those with score above 0.001 amount to about 10% of the whole entries. Thus, we exclusively focus on entries scoring above 0.01. We find that this value is the balance between precision and speed. The reduction of this critical value means more entries needed to be labelled by human. The entries labelled by human in Fig. 2 include two stages. The first one is to judge whether

the entry belongs to the target field or not, then it is to label the entries with key words based on the personal knowledge. Therefore, some of the manually labelled entries situated in the ambiguity region as shown in Fig. 2 do not belong to the target field but are labelled by other key words describing other fields. This ambiguity region holds vital significance for further research, which we will discuss later.

After approximately 30–40 cycles of updating the model, we achieve a robust model for classifying the literatures, and simultaneously obtain a dataset for this model having about 10k labelled entries (200–300 entries in each cycle labelled by both human and the model). Throughout this process, only around 1000–4000 entries are labelled by human. If we were to expand this resultant dataset, for example, to include 1M entries, the number of entries labelled by human might be less than 10k–15k since only about 1%–3% of the whole entries score between 0.01 and 0.9 as we find in 'battery' model.

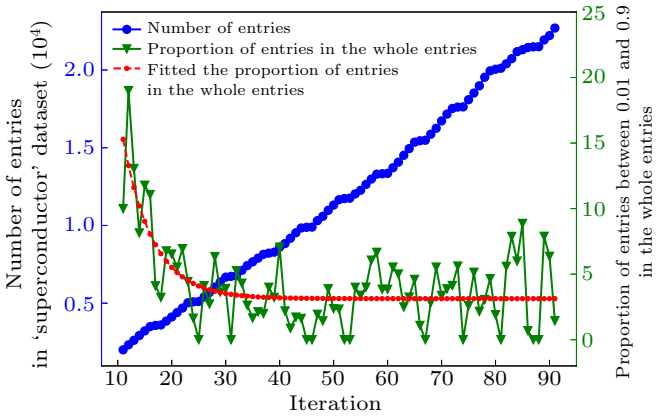



**Fig. 3.** The number of entries in 'superconductor' dataset and the percentage of the entries which score between 0.01 and 0.9 (i.e., the ambiguity region) changes with the iteration cycles when creating the 'superconductor' model.

The detail for training the 'superconductor' model is presented in Fig. 3. In the first few cycles, it gave quite bad results with major entries falling into the ambiguity region. It became better after about seven cycles. Therefore, we show the details from the $10^{\text{th}}$ cycle with about 1023 entries labelled. We observed an emergent distinction during the 10 to 20 iterations, namely, the percentage of the ambiguity entries reduces to about 5% as shown in Fig. 3. After about 25 cycles, the model became robust. All data in Fig. 3 is counted based on the setting of ambiguity region as those with score between 0.01 and 0.9. In fact, after about 20–30 cycles, this setting (0.01 and 0.9) could be adjusted since the model has been improved much and the number of entries to be labelled by human may be further reduced.

## 4. Data analysis

According to the aforementioned flow chart, we possess approximately 90k labelled entries in 'battery' dataset, 25k labelled entries in 'superconductor' dataset, 15k labelled entries in 'topological material' dataset, and 70k labelled entries in 'AI in materials science' dataset. At the same time, the ambiguity dataset holds substantial importance for NLP models. These entries offer an opportunity to reveal potential relationships among different fields.

The key words in the ambiguity datasets from the above four models are presented in Table 1. There are several key words describing various research topics in each model as labelled manually, and some of these key words might appear in more than two models. They in fact demonstrate the connections between different research fields. For example, superconductor, topological materials, and quantum computing appear in two models of 'superconductor' and 'topological material' since the topological superconductor realized in topological material is highly expected for topological quantum computing. These relationships occur naturally as researchers in these fields communicate with each other and their literatures inherently exhibit this connection. Similarly, we observe analogous phenomena in the 'battery' and 'AI in materials science' models.

**Table 1.** The selected key words in ambiguity dataset for four models.

| Models | Key words in ambiguity dataset |
|---|---|
| Battery | energy such as catalysis, solar energy, supercapacitor |
| Superconductor | topological materials, magnetic materials, quantum computing |
| Topological material | superconductors, quantum computing |
| AI in materials science | DFT, AI in drug, numerical simulation |

## 5. Model prediction power

To conduct a more in-depth analysis, we selected another totally different dataset (different from the dataset used in training the model) of 150k entries from Web of Science[30] for testing the prediction power of the model, which primarily contains the abstracts of various literatures, with a focus on journals of physical and materials sciences. The score distribution of these entries is depicted in Fig. 4. Clearly, most of the entries have scores below 0.01, which means that the literatures classified into a target field are a very small part of the whole dataset. This further confirms the viability and effectiveness of our methods in considering so many subdivisions included in these literatures.

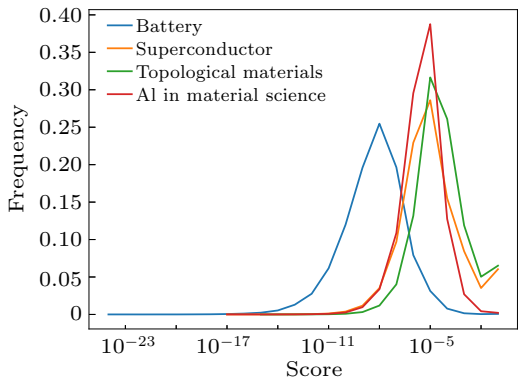



**Fig. 4.** The score distributions for four models.

The research field of battery encompasses complex connections with physics, chemistry, and materials, its relationship with other fields is also intricate. In light of this, we have chosen the 'battery' model for further analysis.

As shown in Fig. 5, there are 6428 entries whose scores exceed $10^{-5}$. Then, we define the confusion matrix elements for them in Table 2. They are positive (negative) if their scores are above (below) 0.5. We find 86 positive entries and 6342 negative ones. The true positive (TP) instances, which score above 0.5 and belong to battery literatures, amount to 82 entries, and the false positive (FP) instances, which score above 0.5 but do not belong to battery literatures, are 4 entries. Further, there are 54 false negative (FN) instances, which score below 0.5 but indeed belong to battery literatures. The left 6288 are true negative (TN) instances, which score below 0.5 and do not belong to battery literatures. The precision of the 'battery' model is 0.953 (82/86) and the recall is 0.603 (82/136). Therefore, the F1 score is calculated to be 0.738, indicating a high level of accuracy and performance for the 'battery' model. It should emphasize that there exist some entries including battery management system (BMS) and Li–S battery based on electrocatalyst, which may not be labelled as 'battery' literature by some researchers, so the recall is underestimated.

We conducted a further analysis of the ambiguity dataset from 'battery' model, including the FP and TN datasets, and the outcomes are depicted in Fig. 6. The center section of Fig. 6 shows the components and research methods in battery and the right section is the closely related content to the center section. Several relationships are evident in the right section, such as those associated with diffusion or superionic conductors. For instance, diffusion is a prevalent phenomenon across physics, chemistry, and biology, while synthesis and dendrite processes are not only linked to diffusion but also commonly found in the anode especially Li. Another notable example is superionic conductors, as solid-state batteries require solid state electrolytes with high ionic conductivity or in a superionic state. Interestingly, this analysis has led to the discovery of a new superionic material derived from the synthesis process.[31]

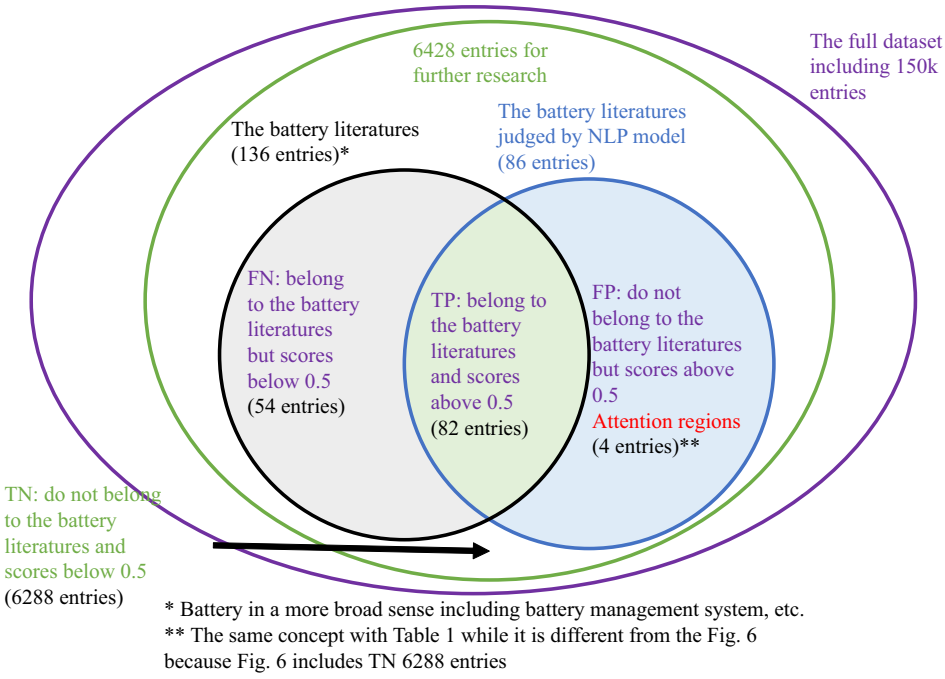



**Fig. 5.** The number of entries in full dataset, those with score above $10^{-5}$, those classified as TP, TN, FN and FP.

**Table 2.** The definition of confusion matrix elements true positive (TP), false positive (FP), false negative (FN) and true negative (TN).

| | The entries belonging to this field | The entries not belonging to this field |
|---|---|---|
| The entries score above 0.5 | true positive (TP) | false positive (FP) |
| The entries score below 0.5 | false negative (FN) | true negative (TN) |

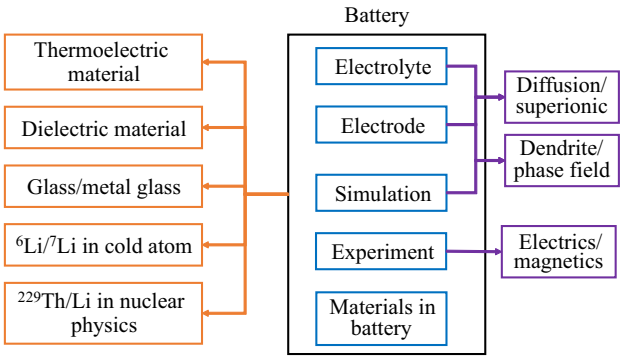



**Fig. 6.** The relationship between battery and other fields given by 'battery' model. Blue frames represent the key words like compositions and methods in battery field. The purple frames represent some known relationship with other fields and the orange frames represent the unknown relationship with battery revealed by this model.

Among these findings, some unexpected relationships emerged, and they are highlighted in the left section in Fig. 6. The relationship between dielectric materials and the superionic state is revealed and has been illustrated by the discovery of $Li_{0.5}La_{0.5}TiO_3$.[32] Bai *et al.* shed light on the possibility of glass being in a superionic state.[33] Another surprising discovery relates to $^{6}$Li/$^{7}$Li and $^{229}$Th in nuclear physics, which could be attributed to Li and neutron diffusion, for example, Li *et al.* predicted superionic state in $LiH_2$ during nuclear fusion.[34] Such kind of exploration of ambiguity dataset allows us to gain valuable insights into various interdisciplinary connections and correlations. These unforeseen connections offer intriguing insights and open up new avenues for further exploration and research.

Furthermore, we conducted an analysis of the battery dataset present in these literatures (6k battery entries in battery dataset with 90k labelled entries). Specifically, we focused on key physical parameters such as energy density, capacity, electrical conductivity, and diffusion coefficient, which are of primary interest in the field of batteries. The statistical frequencies of the energy density, capacity, electrical conductivity and diffusion coefficient in literatures published between August 2022 and July 2023 are presented in Fig. 7. It is obvious that energy density of battery mainly locates around 200–300 Wh/kg while the capacity is within 100–200 mAh/g. More details are shown in supplementary materials.

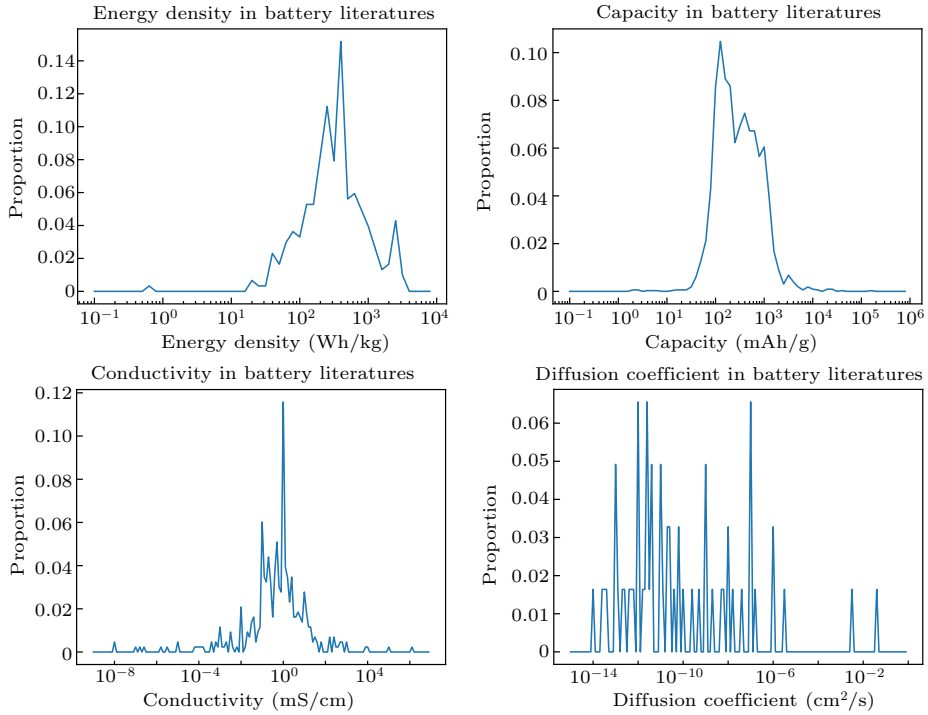



**Fig. 7.** The statistical frequencies of the energy density, capacity, electrical conductivity and diffusion coefficient in literatures published between August 2022 and July 2023.

## 6. Discussion and conclusion

In this article, we introduce a robust and efficient method for constructing the datasets and NLP models dedicated to classifying literatures on condensed matter physics and materials science. Our approach possesses two significant advantages. Firstly, it will reduce the number of human labelled entries substantially by about 90%. Secondly, the ambiguity dataset obtained through this method may uncover potential relationships among different research fields that were unnoticed before. These advantages underscore the effectiveness and great potential of this method in advancing the scientific research in various fields.

The most significant challenge faced by the application of NLP in science is the scarcity of sufficient dataset. However, our method offers a viable solution, providing ample entries to overcome this obstacle. Our NLP code can be ac-

cessed on GitHub or Gitee. We believe that implementing this method will accelerate the advancements of NLP applied in condensed matter physics and materials science. In the near future, we aim to create personalized NLP models based on these entries,[35] tailored to meet the individual needs of researchers. Such personalized approach will further enhance the efficiency and effectiveness of NLP applications.

## Program availability

The code has been published on GitHub, Gitee, and Science Data Bank. We train it on NVIDIA GeForce RTX 2050 (72 second each epoch for 90k battery entries) and Intel i7-12700F (550 second each epoch for 90k battery entries). The personalized model has been internal testing at the Electronic Laboratory for Materials Science since October 2023.[35] The data about energy density and ionic conductivity will be presented at the website of Data Center for Condensed Matter Physics, Chinese Academy of Sciences[36] step by step at the end of 2023.

## Acknowledgements

The authors thank Prof. Baohua Zhang in Computer Network Information Center (CNIC) of the Chinese Academy of Sciences (CAS) for building the website.[36]

The authors used the 'AI in materials science' models (v2023.12) to assist the analyze the literatures from 2023.8 to 2024.01 in arXiv which verified that our models are powerful and useful. The authors also use ChatGPT 3.5 (model 0613) to polish the article (based on Microsoft Azure OpenAI service and the prompt is "polish this sentence: …", see more information at https://azure.microsoft.com/zh-cn/products/cognitive-services/openai-service/) but the authors ensure it doesn't change the original intention.

This research was funded by the Informatization Plan of Chinese Academy of Sciences (Grant No. CAS-WX2021SF-0102), the National Key R&D Program of China (Grant Nos. 2022YFA1603903, 2022YFA1403800, and 2021YFA0718700), the National Natural Science Foundation of China (Grant Nos. 11925408, 11921004, and 12188101), and the Chinese Academy of Sciences (Grant No. XDB33000000).

## References

[1] Tshitoyan V, Dagdelen J, Weston L, Dunn A, Rong Z, Kononova O, Persson K A, Ceder G and Jain A 2019 *Nature* **571** 95
[2] Swain M C and Cole J M 2016 *J. Chem. Inf. Model.* **56** 1894
[3] Devlin J, Chang M W, K Lee and Toutanova K 2018 *arXiv* 1810.04805
[4] Brown T B, Mann B, Ryder N, *et al.* 2020 arXiv 2005.14165
[5] Song Y, Miret S and Liu B 2023 arXiv 2305.08264
[6] Liu Y, Yang Z W, Zou X X, Ma S C, Liu D H, Avdeev M and Shi S Q 2023 *National Science Review* **10** nwad125
[7] Mikolov T, Chen K, Corrado G and Dean J 2013 arXiv 1301.3781
[8] Levy O, Goldberg Y and Dagan I 2015 *Transactions of the Association for Computational Linguistics* **3** 211
[9] Schnabel T, Labutov I, Mimno D and Joachims T 2015 *Conference on Empirical Methods in Natural Language Processing* 298–307
[10] Arora S, Li Y Z, Liang Y Y, Ma T Y and Risteski A 2016 *Transactions of the Association for Computational Linguistics* **4** 385
[11] Yin, Zi, Shen and Yuanyuan 2018 arXiv 1812.04224
[12] Hopfield J J 1982 *Proc. Natl. Acad. Sci. USA* **79** 2554
[13] Hochreiter S and Schmidhuber J 1997 *Neural Computation* **9** 1735
[14] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser L and Polosukhin I 2017 arXiv 1706.03762
[15] Park Y J, Jerng S E, Park J S, Kwon C, Hsu C W, Ren Z, Yoon S and Li J 2023 arXiv 2308.13687
[16] Yang S J, Li S, Venugopalan S, Tshitoyan V, Aykol M, Merchant A, Cubuk E D and Cheon G 2023 arXiv 2311.13778
[17] Zhang X, Zhou Z C, Chen M and Sun Y Y 2023 *J. Phys. Chem. Lett.* **14** 11342
[18] Rubungo A N, Arnold C, Rand B P and Dieng A B 2023 arXiv 2310.14029
[19] Sagawa T and Kojima R 2023 arXiv 2311.06708
[20] Buehler M J 2023 arXiv 2310.19998
[21] Zheng Z, Zhang O, Borgs C, Chayes J T and Yaghi O M 2023 *J. Am. Chem. Soc.* **145** 18048
[22] Zheng Z L, Alawadhi A H, Chheda S, Neumann S E, Rampal N, Liu S C, Nguyen H L, Lin Y, Rong Z, Siepmann J I, Gagliardi L, Anandkumar A, Borgs C, Chayes J T and Yaghi O M 2023 *J. Am. Chem. Soc.* **145** 28284
[23] Zheng Z L, Rong Z C, Rampal N, Borgs C, Chayes J T and Yaghi O M 2023 *Angewandte Chemie* **135** e202311983
[24] Zheng Z L, Zhang O F, Nguyen H L, Rampal N, Alawadhi A H, Rong Z C, Head-Gordon T, Borgs C, Chayes J T and Yaghi O M 2023 *ACS Cent. Sci.* **9** 2161
[25] Boiko D A, MacKnight R, Kline B and Gomes G 2023 *Nature* **624** 570
[26] Yang X J, Wilson S D and Petzold L 2024 *arXiv* 2401.01089
[27] Chen Z Y, Xie F K, Wan M, Yuan Y, Liu M, Wang Z G, Meng S and Wang Y G 2023 *Chin. Phys. B* **32** 118104
[28] Jyotirmoy Deb, Lakshi Saikia, Kripa Dristi Dihingia, G. Narahari Sastry 2024 *J. Chem. Inf. Model.* **64** 799
[29] TensorFlow https://tensorflow.google.cn/
[30] Web of Science https://www.webofscience.com/wos/woscc/basic-search
[31] Wu S Y, Xiao R J, Li H and Chen L Q 2023 arXiv 2304.08728
[32] Inaguma Y, Chen L Q, Itoh M, Nakamura,T Uchida T, Ikuta H and Wakihara M 1999 *Solid State Commun.* **86** 689
[33] Chang C, Zhang H P, Zhao R, Li F C, Luo P, Li M Z and Bai H Y 2022 *Nat. Mater.* **21** 1240
[34] Li F, Wang H, Li J and Geng H 2023 *Chin. Phys. B* **32** 106103
[35] The personalized model has been internal testing at Materials Science Electronic Laboratory Platform https://in.iphy.ac.cn/eln/#/recusertype
[36] Data Center for Condensed Matter Physics, Chinese Academy of Sciences http://condmatt.iphy.ac.cn/. The model results have been presented weekly at http://condmatt.iphy.ac.cn/literature_trends/newest and the data about energy density and ionic conductivity will be merged in it step by step. The authors claim that the model has been changed from TensorFlow to pyTorch based on transformers architecture since August 2023 at http://condmatt.iphy.ac.cn/literature_trends/newest. This website will be discard and your can see [35] or wait for new website.