J Vis (2024) 27:1213–1226
https://doi.org/10.1007/s12650-024-01010-z

REGULAR PAPER

Hongjiang Lv · Zhibin Niu · Wei Han · Xiang Li

Can GPT embeddings enhance visual exploration of literature datasets? A case study on isostatic pressing research

Received: 21 June 2023 / Revised: 31 January 2024 / Accepted: 14 May 2024 / Published online: 7 June 2024

Abstract Visual exploration of literature datasets, especially in specialized domains like isostatic pressing in materials research, aids scientific understanding and discovery but demands robust natural language processing techniques for semantic representation. Existing methods often rely on complex and time-consuming processes to obtain text embeddings, which are numerical representations of text that capture their semantic information and similarity. The quality of text embeddings is crucial for enabling visual exploration of literature datasets. Our research question is whether visual exploration of literature datasets can benefit from GPT (generative pre-trained transformer) text embeddings. We seek to answer this question by performing case studies and expert interviews. To do this, we curated a unique literature dataset about isostatic pressing, sourced from diverse periods and genres. Utilizing a GPT embedding model, we generated embeddings for textual analysis, visualizing and examining their semantic interrelations. Expert reviews were undertaken to evaluate the utility of these techniques. Our findings show that GPT text embeddings offer significant improvements in visually exploring literature datasets, revealing deep semantic similarities and diversities. We also discuss the implications, limitations of our study, and propose directions for future research.

Keywords Visualization and visual analytics · Generative pre-trained transformer (GPT) · Text embeddings · Isostatic pressing · Powder metallurgy · Materials informatics

1 Introduction

Visual exploration of literature datasets is a challenging task that requires natural language processing techniques to extract and represent the semantic information and diversity of texts from different domains

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12650-024-01010-z.

H. Lv · Z. Niu (✉)
College of Intelligence and Computing, Tianjin University, Yaguan Road No.135, Tianjin 300354, China
E-mail: zniu@tju.edu.cn

H. Lv
E-mail: lvhongjiang@tju.edu.cn

W. Han · X. Li
China Iron and Steel Research Institute Group, Beijing, China
E-mail: hanw@cisri.com.cn

X. Li (✉)
University of California, Berkeley, Berkeley, CA, USA
E-mail: xli@berkeley.edu

and genres. Literature datasets contain rich and complex texts that can span across multiple topics, styles, perspectives, and time periods. Analyzing and understanding such texts can provide valuable insights for researchers, educators, students, and readers.

One of the key techniques for visual exploration of literature datasets is text embeddings (Federico et al. 2017), which are numerical representations of text that capture the semantic meaning and similarity of words, sentences, or documents. Text embeddings can be used to map high-dimensional text data into low-dimensional spaces that can be easily visualized and compared using various methods such as graphs, charts, or tables. Text embeddings can also enable various tasks such as clustering, classification, topic modeling, or anomaly detection on literature datasets. However, not all text embeddings are equally suitable for visual exploration of literature datasets. Existing methods (Levy and Goldberg 2014; Liu et al. 2018; Peters et al. 2018; Jing and Xu 2019; Yang et al. 2019) can not capture the nuances and terminology of specific domains or genres. Moreover, some text embeddings may not preserve the global structure or local details of the original text data, which can affect the quality and relevance of the visualizations (Liu et al. 2019; Alharbi and Laramee 2019).

ChatGPT−3.5/4 by OpenAI has drawn significant attention these days. Its powerful embedding technique can potentially enhance the visualization of literature data. In this paper, we attempt to assess OpenAI's GPT text embeddings, pre-trained on large and diverse corpora, for visual exploration of literature datasets. These embeddings are powerful, compact, expressive, and easy to use. Our approach leverages the OpenAI API to directly utilize the sophisticated GPT models for generating text embeddings, which circumvents the complexities of traditional NLP training methods and capitalizes on the enhanced representational power of GPT embeddings for our visualization system. We use a case study on isostatic pressing research, a domain-specific literature dataset with texts from different sources and time periods. Isostatic pressing is an advanced manufacturing process that applied uniform pressure over the entire produce to increase its density and reduce internal defects. We work closely with isostatic pressing experts to collect, preprocess, evaluate, and validate our data and results. We use an OpenAI embedding model to get vectors for our texts, and visualize and analyze them with t-SNE (Maaten and Hinton 2008), UMAP (McInnes and Healy 2018), and BERTopic (Grootendorst 2022). We conduct expert interviews and compare our embeddings with "state-of-the-art" alternatives.

To the best of our knowledge, our work is the first to evaluate GPT text embeddings on enhancing the visual exploration of literature datasets. We demonstrate the effectiveness and usefulness of it through a case study on the isostatic pressing dataset and validated with experts. We show that GPT text embeddings can provide a comprehensive and intuitive visual exploration of literature datasets, and highlight the semantic similarity and diversity of literature texts. We also conduct an experts case and user study to evaluate our tool and compare it with existing tools. We discuss the implications and limitations of our approach, and suggest directions for future work.

2 Related work

Literature dataset visualization Visual exploration of literature datasets is a challenging and important task that can help researchers, educators, students, and practitioners discover and understand relevant literature in their fields. Several tools have been developed to support this task, such as CiteSpace (Chen 2006), ConnectedPapers, VOSviewer (Eck and Waltman 2009, CitNetExplorer (Eck and Waltman 2014), Galex (Li et al. 2020), and LitMaps (2024). These tools typically use bibliographic data, such as citations, authors, keywords, and abstracts, to generate visualizations of literature networks, clusters, or trends. However, these tools have some limitations, such as relying on small or outdated datasets, lacking support for personalized exploration, or requiring installation or registration.

With the establishment of many literature databases [e.g., MAG (Sinha et al. 2015), AMiner (Tang et al. 2008), Semantic Scholar (Ammar et al. 2008)], it's becoming easier to obtain large amounts of literature data. Plenty of previous studies that used citation networks to analyze the structure and evolution of scientific literature. Garfield (1979) proposed to obtain the law of research evolution by analyzing the temporal distribution of citations in scientific literature. Price (1965) proposed citation network to analyze research trend. Hummon and Dereian (1989) proposed main path analysis method to identify the most important citation path in citation network. Inspired by this idea, we conducted citation path for each literature, it can help experts evaluation the effectiveness of our approach more accurately. There are different approaches to build a reference network, such as bibliographic coupling (Kessler 1963) and co-citation network (Small 1973).

Embeddings for Text Representation Text embeddings are numerical representations of texts that capture their semantic information and similarity. They have many applications in natural language processing, such as search, clustering, recommendation, classification, and anomaly detection. Text embeddings can be generated by different techniques and models, which can be divided into two main categories: (1) Shallow feature Learning, such as LDA model (Blei et al. 2001) and TF-IDF (Ramos 2003). LDA model is a generative topic model to find latent topics in a text corpus. TF-IDF is a statistical method to evaluate the importance of words to documents. In this paper, we use this method to extract keywords for each literature. (2) Deep semantic feature learning based on deep neural network, such as Word2vec (Mikolov et al. 2013, 2013) and Doc2vec (Le and Mikolov 2014). Word2vec uses a skip-gram or continuous bag-of-words architecture to predict the surrounding words of a given word or vice versa. Word2vec embeddings can capture the syntactic and semantic relationships between words, such as similarity, analogy, and compositionality. Doc2vec is an extension of Word2vec that learns document embeddings from large corpora of text. Word2vec and Doc2vec are useful, but they can only model static semantic features. To overcome this shortcoming, many dynamic semantics feature learning models are proposed, such as ELMo (Peters et al. 2018), BERT (Devlin et al. 2019), and GPTs (Radford and Narasimhan 2018; Radford et al. 2019; Brown et al. 2020; Ouyang et al. 2022). ELMo captures dynamic semantic features using a biLSTM model. BERT is a transformer-based model that learns contextualized word and sentence embeddings from large corpora of text. BERT embeddings capture the contextual and semantic information of texts, such as polysemy, coreference, and entailment. Owing to the substantial parameter count of the BERT model, various efforts have been made to develop streamlined variants, such as ALBERT (Lan et al. 2019), DistilBERT (Sanh 2019), TinyBERT (Jiao 2020), and MiniLMv2 (Wang et al. 2021). MiniLMv2 is a knowledge distillation model, which is distilled from different BERT models. “all-MiniLM-L6-v2” is a “state of the art” model based on MiniLMv2, and it has competitive performance Wang et al. (2021), so we use this model as the baseline in our work. Our approach uses a text embedding model named “text-embedding-ada-002” based GPT−3.5 to capture semantic information for each literature.

3 Our approach

3.1 Overview

We now present our method for creating and visualizing the literature dataset with GPT embeddings. We first collected the literature data from various sources to cover a wide range of topics. We used the OpenAI GPT API to get the embeddings for the data, which capture the semantic meaning and context of the texts. We applied t-SNE and UMAP to reduce the dimensionality of the embeddings and plot the data in a scatter graph. We then extracted topics from the data using BERTopic, which clusters the embeddings into coherent groups. We created a dynamic network graph of the citation relations to show the connections and influences among the literature data.

3.2 Data acquisition and data cleaning

We have chosen to use the topic of *isostatic pressing in material research* as a case study for our visual exploration of literature datasets. Isostatic pressing is a suitable topic for our study because it is a domain-specific literature dataset that contains rich and diverse texts from different sources and time periods (e.g., Henderson et al. 2000; Kim and Oh 2009; Ruttert et al. 2016; Li 2021. The texts span across multiple topics, styles, and perspectives. Analyzing and understanding such texts can provide valuable insights for researchers, educators, students, and practitioners in this field. *Moreover, we have worked closely with isostatic pressing experts to collect, preprocess, evaluate, and validate our data and results.*

We created the literature dataset on isostatic pressing research and searched all literature up to July 2022. We searched for literature related to this topic using the keywords “isostatic pressing” in OpenAlex and Scopus, and merged the list of literature found on both platforms. We collected 2365 lature in total, but we filtered out those that were not in English, not peer-reviewed, or had missing key information (e.g., title) metadata. We also eliminated duplicate literature that appeared in multiple sources. We ended up with 1378 lature for our literature dataset. Among them, 367 lature lacked abstract information, which we

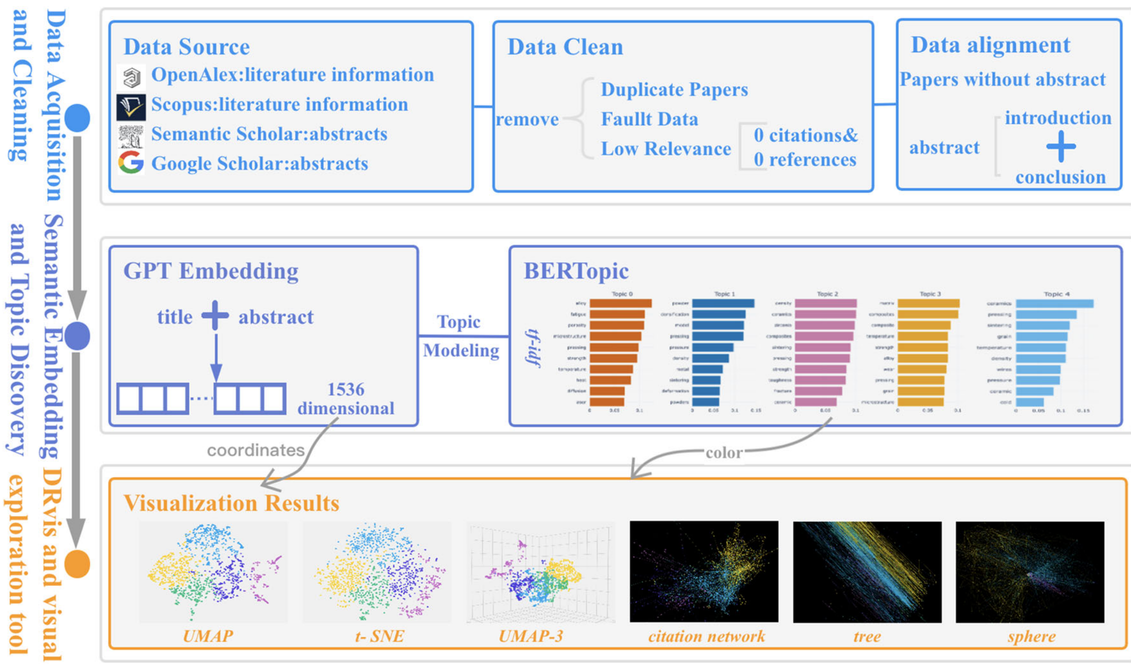

Fig. 1 Overview of the approach, including data preprocessing, embedding pipeline, and visualization

supplemented through Semantic Scholar and Google Scholar. And for the 242 lature that did not have abstracts, we used their introductions and conclusions as substitutes. We then extracted and processed the metadata of each literature, such as title, abstract, authors, keywords, year, and citations, and stored them in a JSON file for further analysis.

3.3 Semantic embedding with GPT

We used the OpenAI GPT API [1] to get embeddings for the literature data. The API provides a powerful model, text-embedding-ada-002, that can generate vector representations of text or code. It is based on GPT−3.5, a "state of the art" to date model that can understand and generate natural language. We chose this model because it can measure the relatedness of text strings and produce high-quality embeddings that capture the semantic meaning and similarity of text data.

We combined the title and abstract of each paper using a distinct separator token, then passed them to the API endpoint, specifying the embedding model 'text-embedding-ada-002' through its unique model ID[2] (Fig. 2). The API returned an embedding vector of 1536 dimensions for each paper. We saved the embedding vectors in a CSV file for further analysis.

3.4 Topic discovery in literature using BERTopic and GPT embeddings

We used BERTopic to analyze the literature and find topics. BERTopic is a topic modeling technique that uses BERT embeddings and clustering algorithms to create dense clusters of texts with similar topics.

We fed the embeddings from GPT to BERTopic to get the topic and label of each paper. We used k-means from scikit-learn (Pedregosa et al. 2011) as the clustering algorithm for BERTopic, and the elbow method (Marutho et al. 2018) to find the optimal number of topics, which was 5. We extracted ten topic words with the highest frequency per topic, as shown as Fig. 3.

The statistical information for each topic is displayed in Table 1, where the keywords are the terms selected by GPT to describe each topic, and the number of literature refers to the quantity of literature belonging to that topic. Then, our cooperated material experts use GPT to generate the main themes for each type based on these keywords. Subsequently, experts obtain the quality of literature classification based on

[1] https://platform.openai.com/docs/guides/embeddings/what-are-embeddings.

[2] https://platform.openai.com/docs/guides/embeddings/use-cases.

```
{
  "data": [
    {
      "input": "Title: Hot isostatic pressing diagrams : new developments; Content:
      The equations and procedures for constructing hot-isostatic pressing diagrams
      are greatly simplified and clarified. In earlier work...",
      "module": "text-embedding-ada-002"
    },
    {
      "input": "Title: Fundamental aspects of hot isostatic pressing: An overview;
      Content: Hot isostatic pressing (hipping) can be used for upgrading castings,
      densifying presintered components, consolidating powders, and interfacial
      bonding. It involves...",
      "module": "text-embedding-ada-002"
    },
    ...
  ]
}
```

Fig. 2 The example of embedding training

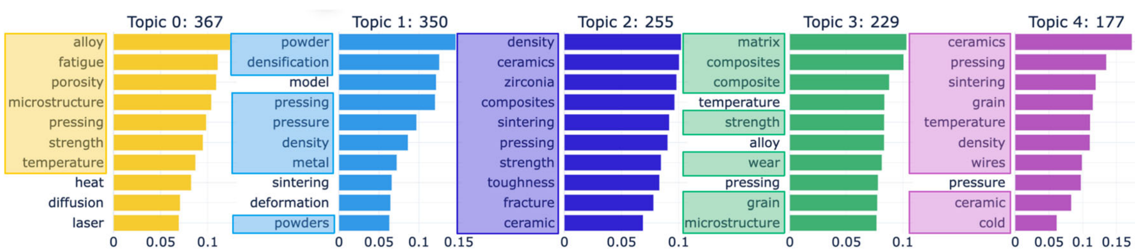

Fig. 3 The bar charts represents the statistical information of the topic words extracted by Bertopic. Each bar chart represents a topic, with the number following the title indicating the number of literature belonging to that topic. The vertical axis displays the ten topic words with the highest frequency, while the horizontal axis shows the frequency of the topic words. The boxed topic words are those included in the topic description sentences generated using GPT for each topic

Table 1 Keywords and statics of literature

Topic ID	Keywords	Number of Literature
Topic0	Alloy, fatigue, porosity, microstructure, pressing, strength, temperature	367
Topic1	Powder, densification, pressing, pressure, density, metal, powders	350
Topic2	Density, ceramics, zirconia, composites, sintering, pressing, strength, toughness, fracture, ceramic	255
Topic3	Matrix, composites, composite, strength, wear, grain, microstructure	229
Topic4	Ceramics, pressing, sintering, grain, temperature, density, wires, ceramic, cold	177

topics. GPT and BERTopic can capture the semantic meaning and context of the texts. This helps us discover relevant and coherent topics that reflect the research trends and challenges in the literature.

3.5 Semantic visualization through dimensionality reduction

We employed t-SNE (Maaten and Hinton 2008) and UMAP (McInnes and Healy 2018) for high-dimensional data visualization, recognizing that initial conditions and hyperparameters critically impact the resulting embeddings. The two techniques diverge primarily due to their distinct loss functions: t-SNE utilizes Kullback–Leibler divergence, while UMAP employs Cross-Entropy, and in their capacity to maintain global data structure, with UMAP generally believed outperforming t-SNE in this regard.

For t-SNE, we focused on optimizing two key hyperparameters: *perplexity* and *learning rate*. Perplexity impacts the balance between local and global aspects of the data, acting as a measure of the effective number of neighbors. Following the suggestions by Van der Maaten and Hinton, we experimented with a range of perplexity values from 5 to 50. The learning rate, which affects the optimization convergence, was set according to (Belkina et al. 2019)'s recommendation of using n/12, where n is the dataset size. We systematically explored learning rates from 10 to 200. Through comparative analysis of the resulting visualizations, we selected the best-performing combination of these parameters.

For UMAP, the two primary hyperparameters are *n_neighbors* and *min_dist*. The first one influences the construction of the high-dimensional graph, essentially controlling how UMAP balances local versus global structure. We tested values from 5 to 50. And the second one determines the minimum spacing between data points in the reduced space, for which we assessed a series from 0.01 to 0.5. The optimal settings for UMAP were identified through an experimental approach similar to that used for t-SNE, by evaluating the clarity and separation of clusters in the visualizations.

In both cases, the determination of hyperparameters was an empirical process, guided by the principles of the respective algorithms and adjusted based on the specific characteristics of our dataset. This approach ensured that the dimensionality reduction visualizations were both informative and representative of the underlying data structure.

3.6 The visual exploration tool of hot isostatic pressing literature

To facilitate expert analysis, we developed a visualization system for our literature dataset, as depicted in Fig. 4. The system features a three-dimensional citation network complemented by auxiliary charts, where nodes represent individual works and edges denote citation links. Subsequent sections detail the visualization techniques and interactive functionalities of the system.

Layout We employ various 3D network force-directed visualizations to facilitate observing the research dynamics of this discipline from multiple perspectives.:

(1) *3D force-directed layout* (default, Fig. 5a): Positions nodes based solely on citation relationships, unconstrained by additional factors. This layout intuitively reveals the network's structure, facilitating the observation of interdisciplinary intersections.

(2) *Bottom-up force-directed layout* (Fig. 5b1): The citation network with citations as edges are directed acyclic graphs (DAGs). This layout organizes nodes in layers by citation chronology, with the root node representing the earliest work. This hierarchical arrangement underscores the temporal progression of publications, highlighting the subject's developmental trajectory.

(3) *3D radial force-directed layout* (Fig. 5c): Distributes nodes based on their degree, placing those with fewer connections closer to the center. This model efficiently distinguishes literature with high citation counts, indicative of influential works warranting in-depth analysis.

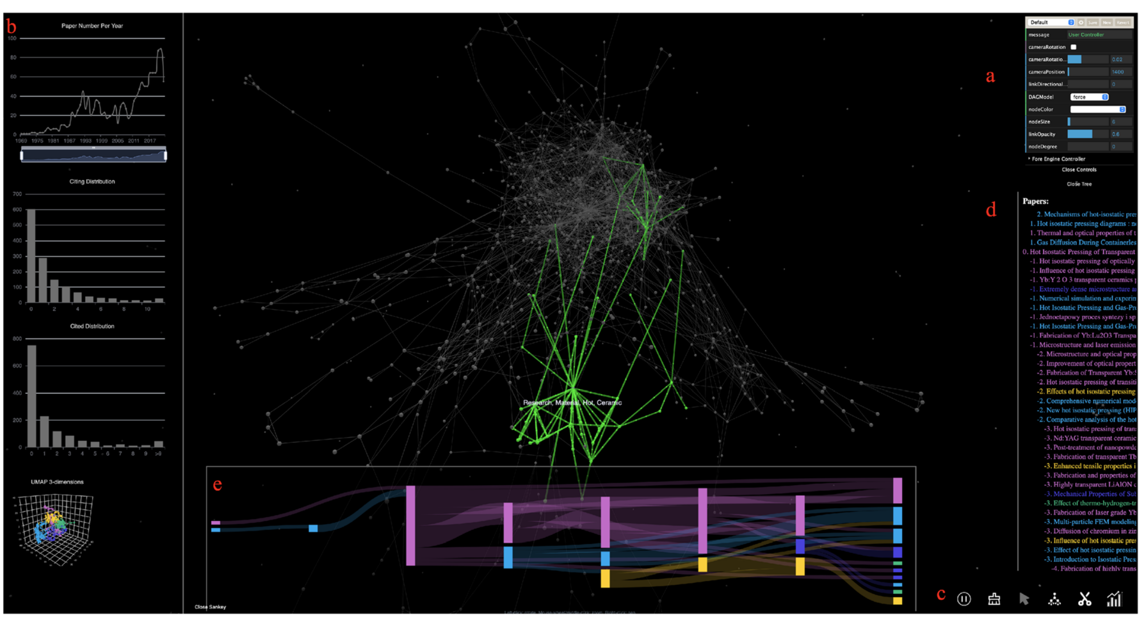

Fig. 4 The interface of our visual exploration tool of Hot Isostatic Pressing literature (http://hip.regvis.net)

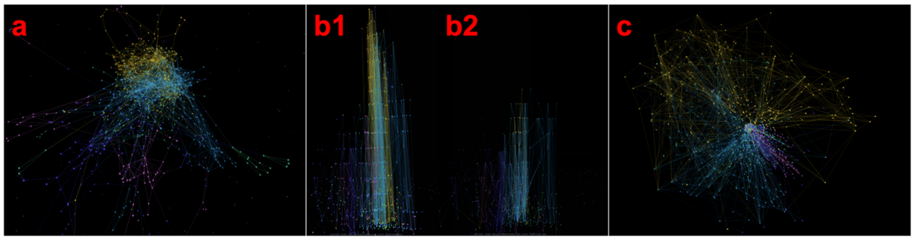

Fig. 5 The citation network is visualized using different layouts, with nodes colored consistently as in Fig. 7. These layouts include: **a** a basic 3D force-directed graph layout, **b1** a bottom-up force-directed layout encompassing all data, **b2** the same bottom-up layout applied to documents published before 2015, excluding papers published after that year, and **c** a 3D radial force layout

Visual encoding strategy Our visual encoding strategy is meticulously crafted to assist users in identifying emergent patterns and evaluating the precision of literature topic labeling derived from various text embedding techniques. In our citation network, colors are utilized to signify the topics to which the literature is associated (Fig. 6). The edges within the network inherit their coloration from the cited literature, ensuring visual coherence; for instance, if literature A cites literature B, the corresponding edge will adopt the color representing literature B's topic. Following extensive consultations with domain experts, we decided against using additional visual attributes (such as node size or shape) to represent factors like citation counts. Such complexity in visual encoding could lead to user confusion and the inadvertent oversight of key information. Instead, we have implemented a suite of interactive features-such as filtering, zooming, and detail-on-demand-which allow users to explore and analyze other literature attributes without overwhelming the primary visual focus on topic identification.

Interaction User interactions with the citation network are categorized into three types: network parameter adjustment, node filtering, and direct network engagement.

- *Network Parameter Adjustment* (Fig. 4a): The control panel allows users to modify network parameters, including node size and edge opacity, to enhance visualization clarity. This functionality aids users in adapting the view for different analytical tasks-larger nodes and reduced edge opacity can obscure the network's structure, while smaller nodes and faint edges may hinder the examination of individual documents.
- *Node Filtering* (Fig. 4b): The statistical charts enable users to filter nodes based on attributes such as publication year, citation count, and more. This feature assists in narrowing down the exploration to specific subsets of the literature.
- *Direct Network Interaction* (Fig. 4c): Users can engage with individual nodes to perform two primary actions: clicking on a node reveals detailed information about the work, including authors, publication

Labeling strategy	Hierarchical paper type		Visual encoding
Experts of Hot Isostatic Pressing manually labeling	Paper Type	Material	
		Manufacturing	
		Simulation	
		Fundamental	
		Review	
	Temperature Type	Hot	
		Warm	
		Cold	
	Material Type	Metal	
		Ceramic	
		Polymer	
		Composite	
		Other INM	
	AM Related	Yes	
Automatically generated paper type	BERTopic generates Type	Topic0	
		Topic1	
		Topic2	
		Topic3	
		Topic4	

Fig. 6 Visual encoding of different labeling strategies

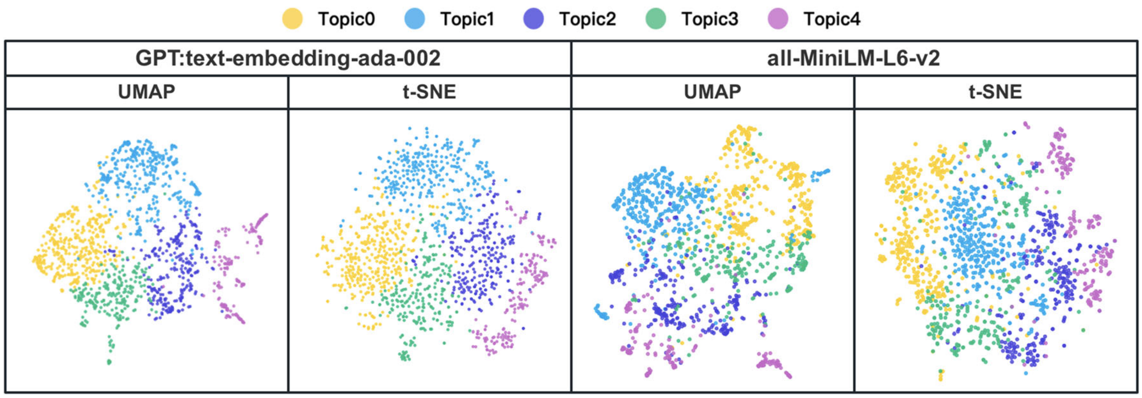

Fig. 7 A 2D projection of the Isostatic Pressing literature dataset using UMAP (n_neighbors=20, min_dist=0.1) and t-SNE (perplexity=30, learning_rate=10). The papers are plotted as points with labels and colors based on their topic

date, and journal; and selecting a node highlights all related literature (as shown in Fig. 4), recursively tracing citations to and from the focal document. Corresponding titles are listed on the page's right side, with links to their OpenAlex detail pages (Fig. 4d). Additionally, a Sankey diagram visualizing citation relationships among the highlighted literature is rendered below (Fig. 4e), clarifying citation flows across different subjects.

4 Evaluation

To answer the question "Can GPT Text Embeddings Enhance Semantic Visual Exploration of Literature Datasets?," we report the visualization results, conduct case studies, and perform expert review.

4.1 Visualization results

We first examined the visualization results of our approach, namely the t-SNE and UMAP plots (Fig. 7) and the citation network (Figs. 4, 5, 8). We evaluated the visualization results of our algorithm against a baseline approach employing "all-MiniLM-L6-v2" text embeddings, as opposed to OpenAI GPT text embeddings.

Each topic has 10 keywords extracted by BERTopic, and then we generated a summary sentence for each topic using GPT. Among them,

Topic 0: focuses on the influence of microstructure, temperature, and pressing process on fatigue behavior, strength, and porosity in alloy-based materials.

Topic 1: focuses on powder metallurgy, exploring the effects of pressing and pressure on densification, deformation, and density of metal powders.

Topic 2: focuses on the properties of ceramic and composite materials, particularly zirconia, examining the impact of pressing, sintering, and fracture characteristics on density, strength, and toughness.

Topic 3: focuses on the study of composite materials, likely metal matrix composites, investigating the influence of pressing on grain structure, microstructure, strength, and wear properties.

Topic 4: focuses on the pressing and sintering of ceramics, studying how these processes affect grain structure, density, and temperature behavior, with potential emphasis on wire applications or cold pressing techniques.

Our method generated more coherent and diverse clusters of papers, demonstrating superior semantic similarity and diversity recognition compared to the baseline. For example, in the t-SNE plot, our approach makes a clear distinction between five topics. In contrast, the baseline approach mixed papers on different topics together, such as Topic 3 (isostatic pressing on metal matrix composites) and Topic 1 (isostatic pressing on metal materials). We then compare the UMAP plot with t-SNE plot. Obviously, UMAP plot showed more fine-grained clusters of papers and sharper clustering boundaries.

4.2 Case studies

Our case studies delve into the advanced functionalities of the system, highlighting its role in the examination of research topic evolution and in the assessment of document labeling techniques. By capturing the nuances of subject development and providing tools for meticulous label evaluation, the system stands as an invaluable resource for experts in the field. The ensuing scenarios encapsulate its application in scrutinizing the "isostatic pressing" domain and in contrasting the outcomes of various labeling algorithms.

Exploring research evolution in isostatic pressing The system's first utility is demonstrated in its capacity to assist experts in mapping the historical development of "isostatic pressing." Utilizing the bottom-up force-directed layout of the citation network, as depicted in Fig. 5b1, documents are chronologically arranged with the seminal works at the foundation. This visual arrangement enables an intuitive understanding of how different research topics have branched out and progressed over time. The inclusion of a "publication year filter" (Fig. 4b) empowers experts to dynamically adjust the citation network to focus on temporal shifts in prominent topics. An illustrative instance (Fig. 5b2) demonstrates that by excluding publications beyond 2015, a noticeable decline in the volume of Topic 0 papers (isostatic pressing on alloy-based materials ●) is observed. This trend signifies a burgeoning interest and impact in alloy-based materials research within the "isostatic pressing" area.

Assessing the precision of document labeling methodologies Beyond tracing topic trajectories, our system enables users to meticulously evaluate and compare the accuracy of document labels derived from various algorithms. Through a side-by-side comparison of node colorings reflecting different labeling methods, domain experts can discern subtle differences and irregularities. The system's interactive features, such as filtering and displaying detailed literature information, facilitate a granular analysis of the labeling accuracy. We replaced the method of obtaining literature embeddings in the approach with MiniLMv2 (Wang et al. 2021) instead of GPT and used it as a baseline method for comparison. In the comparison of citation networks, inconsistencies are easily identified, as shown in Fig. 8 where nodes "c1.1", "c1.2", "c2.1", and "c2.4" display notable discrepancies in labeling. The information of all the literature within the two clusters in Fig. 8 is shown in Table 2. By leveraging the detailed literature information and additional visual aids like "Sankey diagrams" (Fig. 4e), experts are able to corroborate the insufficient precision of the baseline labeling method. For example, the literature "c1.1" pertains to alloy-based materials (Topic0 ●), yet its primary focus is on the impact of rubber molds on the densification behavior of aluminum alloy powders, emphasizing powder metallurgy (Topic1 ●). Our approach assigns it the label topic1, while the

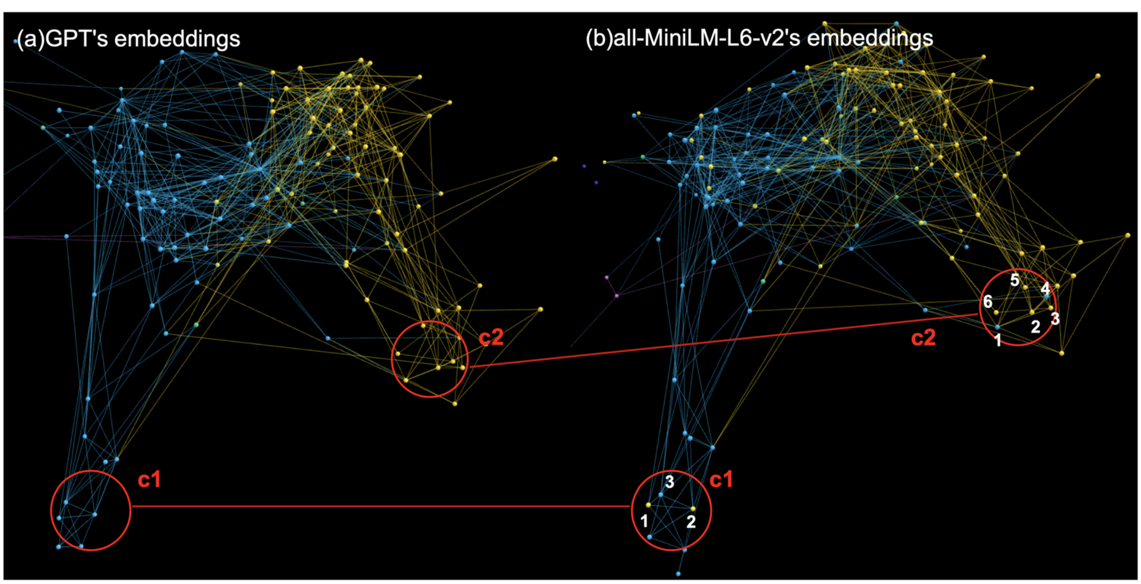

Fig. 8 Part of the citation network, same color encoding with Fig. 7. **a** and **b** are the same subnetwork containing clusters of two topics(denoted by the red circles), namely "c1" and "c2". In **a**, the colors are based on GPT embeddings, whereas in **b**, they are based on baseline model embeddings. For ease of description, we will use the combination of the cluster to which the literature belongs and the cluster's internal numbering (the white numbers next to the points) as the literature ID. For example, "c1.1" refers to the literature with the number 1 within cluster "c1"

baseline method labels it as topic0. A comparable situation is observed with the literature “c1.2”: despite its relevance to alloy-based materials (Topic0), its core subject matter remains powder metallurgy (Topic1). Similarly, within cluster “c2”, there are two literature: “c2.1” and “c2.4”, Both discuss metal powders (Topic1) but center their research on alloy-based materials (Topic0). Expert evaluation of these four literature with discrepant labels has led to the conclusion that our method yields more precise labeling outcomes. The system’s rich interactivity allows for the swift identification and evaluation of “anomalous” nodes, enabling experts to make informed decisions regarding the reliability of different labeling techniques.

4.3 Expert review

We invited two experts, both highly experienced in isostatic pressing research, to participate reviewing the results. Expert 1, a materials science and engineering professor, has published over 30 papers on isostatic pressing, while Expert 2, a senior researcher from a national material laboratory, has dedicated over a decade to the field. We requested these experts to explore the isostatic pressing literature dataset using our approach, and respond to a set of semi-structured questions. These questions aimed to gauge their perspectives on the literature’s quality and relevance, the visualizations’ interpretability and utility, the reliability of the text embeddings and citation network, as well as the strengths and weaknesses of our approach. We also solicited their suggestions. For analysis, we recorded and transcribed their responses.

Table 2 The list of literature in Fig. 8

ID	Literature title	GPT’s label	MiniLMv2’s label
c1.1	Finite element modeling of cold isostatic pressing (Henderson et al. 2000)	Topic1	Topic0
c1.2	Rubber isostatic pressing of metal powder under warm temperatures (Yang et al. 2004)	Topic1	Topic0
c1.3	Rubber isostatic pressing and cold isostatic pressing of metal powder (Yang et al. 2004)	Topic1	Topic1
c2.1	Impact of hot isostatic pressing on microstructures of CMSX-4 Ni-base superalloy fabricated by selective electron beam melting (Ruttert et al. 2016)	Topic0	Topic1
c2.2	Microstructural evolution of creep-induced cavities and casting porosities for a damaged Ni-based superalloy under various hot isostatic pressing conditions (Wang et al. 2015)	Topic0	Topic0
c2.3	Effect of high pressure on the solid-liquid phase change of a nickel base superalloy during hot isostatic pressing (Kim and Oh 2009)	Topic0	Topic0
c2.4	Application research progress of hot isostatic pressing technology in nickel-based single crystal superalloy (Wang et al. 2019)	Topic0	Topic1
c2.5	Effect of the cooling rate during heat treatment and hot isostatic pressing on the microstructure of a SX Ni-superalloy (Lopez-Galileaa et al. 2014)	Topic0	Topic0
c2.6	Influence of hot isostatic pressing on structure and properties of titanium cold-spray deposits (Petrovskiy et al. 2019)	Topic0	Topic0

- How do you evaluate the quality and relevance of the literature dataset on isostatic pressing?
- How do you interpret the visualizations of text embeddings that we have shown you, such as t-SNE, UMAP, and BERTopic? What do they tell you about the semantic similarity and diversity of literature texts on isostatic pressing?
- How accurate and reliable are the visualizations of text embeddings for representing the semantic information and diversity of literature texts on isostatic pressing? Do they capture the nuances and terminology of the domain?
- How accurate and reliable is the visualization of citation network (especially the coloring) for representing the development and citation evolution of isostatic pressing research? Does it capture the trends and patterns of the field?
- Do you have any suggestions or feedback on how to improve the visual exploration of literature datasets on isostatic pressing using text embeddings and citation network?

The experts evaluated the quality and relevance of the isostatic pressing literature we collected, confirming its validity based on various criteria. They gave high ratings for the utility and intuitiveness of the visualizations of text embeddings in literature exploration. The experts found the visualizations easy to interact with, providing a succinct and clear overview of the literature dataset. Their professional domain feedbacks are as follows:

Expert 1: After examining the five topics generated by our approach, Expert 1 commended its innovative categorization strategy for isostatic pressing literature. The proposed method consolidates the traditional, expansive multi-label classification scheme, which typically encompasses labels such as material type, research direction, and manufacturing temperature, into a concise framework of five specific themes.

This classification differentiates primary isostatic pressing research areas: metallic materials (topics 0–1), composite materials (topics 2–3), and less common studies on traditional ceramic materials and cold isostatic pressing (topic 4). It further delineates metallic and composite materials research into detailed topics: the isostatic pressing production process of metallic materials, including design, parameter optimization, and material characterization testing (topic 0); fundamental knowledge in hot isostatic pressing process, such as simulation models and principles derivation (topic 1); metal matrix composites (topic 2); and ceramic matrix composites (topic 3).

Expert 1 recognized this classification strategy's novelty and effectiveness in conjunction with the visualization tool for exploring isostatic pressing research trends. However, he noted that about 5% of the papers were inaccurately grouped, suggesting potential improvements through model enhancements or additional data incorporation, such as paper introductions and conclusions.

Expert 2: Expert 2, while endorsing the effectiveness of our tool's literature grouping and visualization, took interest in the keywords and summary sentences generated for each topic. The expert noted that our approach not only achieved precise topic clustering but also provided accurate summaries, capturing intricate details often missed in human annotation. For example, in topic 4, it linked cold isostatic pressing research with its primary application, ceramic materials. Notably, it identified the keyword "wire," spotlighting a leading research trend, MgB_2 superconducting wires, a prominent research direction in cold isostatic pressing. **Nonetheless**, expert 2 noted that despite the OpenAI text embeddings' powerful semantic understanding ability, its lack of domain-specific knowledge in isostatic pressing resulted in less accurate and specialized terminology in the topic summaries. For instance, in the summary sentence of topic 1, "powder metallurgy" was too generic, as isostatic pressing is only one of many powder metallurgy processes (Li et al. 2023). Similarly, GPT's misinterpretation of "wire" in topic 4 as "wire applications" made it less comprehensible to materials experts. The term "MgB_2 superconducting wires" would be more appropriate. Expert 2 suggested preloading the model with isostatic-pressing-related information to potentially achieve better outcomes.

The experts concluded that, although the OpenAI GPT text embeddings presented a few minor imperfections (as noted above), the overall effectiveness and reliability of the visualizations in showcasing the semantic content and diversity of isostatic pressing literature remained unimpaired. They acknowledged that the text embeddings accurately captured the domain's subtle and specific language, aligning with their own understanding and expertise. They also found the text embeddings to be stable and consistent across various papers and topics. To confirm the effectiveness and trustworthiness of the text embeddings, the experts examined the titles and abstracts of isostatic technical papers grouped by our approach. The domain experts observed that the vast majority of the papers seemed to be appropriately grouped and connected, based on their extensive experience and qualitative assessment of the system's performance.

5 Discussion

In this section, we discuss the main findings and implications of our approach, as well as the limitations and challenges that we faced.

Findings and Implications Our approach has shown that OpenAI GPT text embeddings can enhance the semantic visual exploration of literature datasets. We have demonstrated that OpenAI text embeddings can capture the semantic similarity and diversity of papers in a high-dimensional space, and that they can be projected into a low-dimensional space for interactive visualization using t-SNE and UMAP. We have also shown that OpenAI text embeddings can be combined with citation network to show the development and citation evolution of the literature. Our approach has helped users discover and explore relevant papers in a literature dataset and has provided a clear and concise overview of the literature dataset. Our approach has also been evaluated by domain experts, who have confirmed the quality and performance of our approach, and have provided positive feedback and suggestions.

Limitations and Challenges One limitation is that such an approach relies on the availability and quality of OpenAI text embeddings, which are not always accessible or reliable for all types and domains of texts. For example, OpenAI text embeddings may not be able to handle texts that are too long or too short, or texts that contain specialized or technical terms that are not common in natural language. Another challenge is that our approach does not compare or contrast more different text embedding methods, which may be useful for users who want to see how different text embedding methods perform or differ for visual exploration of literature datasets. We plan to improve this part of the content in our future work.

6 Conclusion

This paper investigates the potential of GPT's text embeddings to facilitate the semantic visual exploration of literature datasets. We use a case study of isostatic pressing papers from 30 years to demonstrate how OpenAI text embeddings can capture the semantic features of the papers in a high-dimensional space. We apply BERTopic to identify the topics and clusters for visual encoding, and compare the visualization outcomes using t-SNE and UMAP. We conduct an expert review to assess the quality of the embedding and visualization techniques. As future work, we aim to expand the study with more diverse and comprehensive literature datasets, and design suitable visual analytics to better compare different text embedding methods.

Data availability The data that support the findings of this study are available upon request from the corresponding author.

Declarations

Conflict of interest The authors have no conflict of interest to disclose.

References

Alharbi M, Laramee RS (2019) SoS TextVis: an extended survey of surveys on text visualization. Computers 8:17

Ammar W, Groeneveld D, Bhagavatula C, Beltagy I, Crawford M, Downey D, Dunkelberger J, Elgohary A, Feldman S, Ha VA, Kinney RM, Kohlmeier S, Lo K, Murray TC, Ooi H-H, Peters ME, Power JL, Skjonsberg S, Wang LL, Wilhelm C, Yuan Z, Zuylen M, Etzioni O (2018) Construction of the literature graph in semantic scholar. ArXiv arXiv:1805.02262

Belkina AC, Ciccolella CO, Anno R, Halpert R, Spidlen J, Snyder-Cappione JE (2019) Automated optimized parameters for t-distributed stochastic neighbor embedding improve visualization and analysis of large datasets. Nat Commun 10(1):5415

Blei DM, Ng A, Jordan MI (2001) Latent Dirichlet Allocation. J Mach Learn Res 3:993–1022

Brown T.B, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T.J, Child R, Ramesh A, Ziegler D.M, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskever I, Amodei D (2020) Language models are few-shot learners. ArXiv arXiv:2005.14165

Chen C (2006) CiteSpace II: detecting and visualizing emerging trends and transient patterns in scientific literature. J Am Soc Inform Sci Technol 57(3):359–377

Devlin J, Chang M.-W, Lee K, Toutanova K (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. ArXiv arXiv:1810.04805

Eck NJ, Waltman L (2009) Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics 84:523–538

Eck NJ, Waltman L (2014) CitNetExplorer: a new software tool for analyzing and visualizing citation networks. J Informet 8:802–823

Federico P, Heimerl F, Koch S, Miksch S (2017) A survey on visual approaches for analyzing scientific literature and patents. IEEE Trans Vis Comput Gr 23:2179–2198

Garfield E (1979) Citation indexing: Its theory and application in science, technology, and humanities, vol 8. Wiley, New York

Grootendorst M (2022) BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv preprint arXiv:2203.05794

Henderson RJ, Chandler HW, Akisanya AR, Barber H, Moriarty B (2000) Finite element modelling of cold isostatic pressing. J Eur Ceram Soc 20:1121–1128. https://doi.org/10.1016/S0955-2219(99)00280-0

Hummon NP, Dereian P (1989) Connectivity in a citation network: the development of DNA theory. Soc Networks 11:39–63

Jiao X, Yin Y, Shang L, Jiang X, Chen X, Li L, Wang F, Liu Q (2020) TinyBERT: Distilling BERT for natural language understanding. In: Findings of the Association for Computational Linguistics: EMNLP 2020. Association for Computational Linguistics, pp. 4163–4174. https://doi.org/10.18653/v1/2020.findings-emnlp.372

Jing K, Xu J (2019) A survey on neural network language models. ArXiv arXiv:1906.03591

Kessler MM (1963) Bibliographic coupling between scientific papers. Am Doc 14:10–25

Kim MT, Oh OY (2009) Effect of high pressure on the solid-liquid phase change of a nickel base superalloy during hot isostatic pressing. J Alloys Compd 477:224–232. https://doi.org/10.1016/j.jallcom.2008.10.117

Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R (2019) ALBERT: A lite BERT for self-supervised learning of language representations. ArXiv arXiv:1909.11942

Le Q, Mikolov T (2014) Distributed representations of sentences and documents. In: Xing EP, Jebara T (eds) Proceedings of the 31st International Conference on Machine Learning. Proceedings of Machine Learning Research, vol 32. PMLR, Bejing, China, pp 1188–1196

Levy O, Goldberg Y (2014) Neural word embedding as implicit matrix factorization. In: Ghahramani Z, Welling M, Cortes C, Lawrence N, Weinberger KQ (eds) Advances in neural information processing systems, vol 27. Curran Associates Inc, Montreal, Canda

Li X, Cui L, Shonkwiler S, McMains S (2023) Automatic characterization of spherical metal powders by microscope image analysis: a parallel computing approach. J Iron Steel Res Int 30(11):2293–2300. https://doi.org/10.1007/s42243-022-00907-z

Li Z, Zhang C, Jia S, Zhang J (2020) Galex: exploring the evolution and intersection of disciplines. IEEE Trans Visual Comput Graphics 26:1182–1192

Li X (2021) Scalable geometric processing techniques with applications in characterizing additively manufactured composites. Doctoral dissertation, University of California, Berkeley

Liu L, Chen J, Fieguth PW, Zhao G, Chellappa R, Pietikäinen M (2018) From BoW to CNN: two decades of texture representation for texture classification. Int J Comput Vision 127:74–109

Liu S, Wang X, Collins CM, Dou W, Ouyang F, El-Assady M, Jiang L, Keim DA (2019) Bridging text visualization and mining: a task-driven survey. IEEE Trans Vis Comput Gr 25:2482–2504

litmaps (2023). https://www.litmaps.com/. Accessed 29 Jan 2024

Lopez-Galileaa I, Huth S, Theisen W (2014) Effect of the cooling rate during heat treatment and hot isostatic pressing on the microstructure of a SX Ni-superalloy. In: MATEC Web of Conferences, vol 14. EDP Sciences, Giens, France, p 13009. https://doi.org/10.1051/matecconf/20141413009

Maaten L, Hinton GE (2008) Visualizing data using t-SNE. J Mach Learn Res 9:2579–2605

Marutho D, Hendra Handaka S, Wijaya E (2018) Muljono: The determination of cluster number at k-mean using elbow method and purity evaluation on headline news. In: 2018 International seminar on application for technology of information and communication, pp 533–538 . https://doi.org/10.1109/ISEMANTIC.2018.8549751

McInnes L, Healy J (2018) UMAP: uniform manifold approximation and projection for dimension reduction. ArXiv arXiv:1802.03426

Mikolov T, Chen K, Corrado G.S, Dean J (2013) Efficient estimation of word representations in vector space. In: International conference on learning representations

Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J (2013) Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems, vol 26, pp 3111–3119

Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C.L, Mishkin P, Zhang C, Agarwal S, Slama K, Ray A, Schulman J, Hilton J, Kelton F, Miller LE, Simens M, Askell A, Welinder P, Christiano PF, Leike J, Lowe RJ (2022) Training language models to follow instructions with human feedback. ArXiv arXiv:2203.02155

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Louppe G, Prettenhofer P, Weiss R, Weiss RJ, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: Machine learning in Python. ArXiv arXiv:1201.0490

Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L (2018) Deep contextualized word representations. In: North American chapter of the association for computational linguistics

Petrovskiy PV, Sova A, Doubenskaia M, Smurov IY (2019) Influence of hot isostatic pressing on structure and properties of titanium cold-spray deposits. Int J Adv Manuf Technol 102:819–827. https://doi.org/10.1007/s00170-018-03233-5

Price DJ (1965) Networks of scientific papers. Science 149(3683):510–5

Radford A, Narasimhan K (2018) Improving language understanding by generative pre-training

Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I (2019) Language models are unsupervised multitask learners

Ramos J E (2003) Using TF-IDF to determine word relevance in document queries. In: Proceedings of the 1st instructional conference on machine learning (2003)

Ruttert B, Ramsperger M, Roncery LM, Lopez-Galilea I, Körner C, Theisen W (2016) Impact of hot isostatic pressing on microstructures of cmsx-4 ni-base superalloy fabricated by selective electron beam melting. Mater Design 110:720–727. https://doi.org/10.1016/j.matdes.2016.08.041

Sanh V, Debut L, Chaumond J, Wolf T (219) DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. ArXiv arXiv:1910.01108

Sinha A, Shen Z, Song Y, Ma H, Eide D, Hsu B-JP, Wang K (2015) An overview of microsoft academic service (MAS) and applications. In: Proceedings of the 24th international conference on world wide web
Small HG (1973) Co-citation in the scientific literature: a new measure of the relationship between two documents. J Am Soc Inf Sci 24:265–269
Tang J, Zhang J, Yao L, Li J.-Z, Zhang L, Su Z (2008) ArnetMiner: extraction and mining of academic social networks. In: Knowledge discovery and data mining
Wang W, Bao H, Huang S, Dong L, Wei F(2021iLMv2: Multi-head self-attention relation distillation for compressing pretrained transformers. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Association for Computational Linguistics, pp 2140–2151. https://doi.org/10.18653/v1/2021.findings-acl.188 . https://aclanthology.org/2021.findings-acl.188
Wang X, Zhou Y, Zhao Z, Zhang Z (2015) Microstructural evolution of creep-induced cavities and casting porosities for a damaged ni-based superalloy under various hot isostatic pressing conditions. Acta Metallurgica Sinica (Eng Lett) 28:628–633. https://doi.org/10.1007/s40195-015-0241-9
Wang J, Jiang Z (2020) Application research progress of hot isostatic pressing technology in nickel-based single crystal superalloy. In: Guo, Y. (ed.) 2019 The 2nd International Symposium on Hydrogen Energy and Energy Technologies(HEET 2019). E3S Web of Conferences, vol. 155. EDP Sciences, Osaka, Janpan. https://doi.org/10.1051/e3sconf/202015501012
Yang H-K, Kim J, Kim K-C (2004) Rubber isostatic pressing and cold isostatic pressing of metal powder. Mater Sci Eng A-Struct Mater Propert Microstruct Process 382:41–49. https://doi.org/10.1016/j.msea.2004.04.056
Yang H, Lee J, Kim K (2004) Rubber isostatic pressing of metal powder under warm temperatures. Powder Technol 139:240–251. https://doi.org/10.1016/j.powtec.2003.01.001
Yang Z, Dai Z, Yang Y, Carbonell J.G, Salakhutdinov R, Le Q V (2019) XLNet: Generalized autoregressive pretraining for language understanding. In: Neural information processing systems