Scientific progress or societal progress? A language model-based classification of the aims of the research in scientific publications

Mengjia Wu*, Gunnar Sivertsen**, Lin Zhang***, Fan Qi*** and Yi Zhang*

*mengjia.wu@uts.edu.au; yi.zhang@uts.edu.au
0000-0003-3956-7808; 0000-0002-7731-0301
Australian Artificial Intelligence Institute, Faculty of Engineering and Information Technology, University of Technology Sydney, Sydney, Australia

** gunnar.sivertsen@nifu.no
0000-0003-1020-3189
Nordic Institute for Studies in Innovation, Research and Education (NIFU), Oslo, Norway

*** linzhang1117@whu.edu.cn; qifan_joan@whu.edu.cn
0000-0003-0526-9677; 0000-0003-1342-3371
Center for Science, Technology & Education Assessment (CSTEA), Wuhan University, Wuhan, China
Center for Studies of Information Resources, School of Information Management, Wuhan University, Wuhan, China

The classification of research by its aims has been a long-term focus in quantitative science studies and R&D statistics. The classical distinction, used by OECD since 1963, is between basic and applied research. In our prior research, we found it useful to distinguish between scientific and societal progress as the two main research objectives in a quantitative analysis of abstracts in scientific publications, which in turn led to developing and testing an automated method for large-scale classification and further analysis. In this study, we conduct a comprehensive evaluation of existing text classification techniques, including traditional text mining models, pre-trained language models, and large language models (LLMs). Our findings demonstrate that fine-tuning domain-specific pre-trained BERT models remains highly competitive even compared to generative LLMs for our task, resulting in a 5-7% accuracy improvement. Through a case study involving 2.3 million scientific articles, we illustrate how the classification of the main aims of research works across diverse subject categories.

1. Introduction

The literature defined science as "an ordered knowledge of natural phenomena or processes, and the interactions among them", and the aim of scientific research is to discover such an order (Hodes, 1974). The scientific community has attempted to classify scientific research with diverse criteria and foci. For example, basic vs. applied research has been used to distinguish scientific literature (Narin et al., 1976), thereby referring to the knowledge type, while some other criteria, such as quantitative vs. qualitative, observational vs. experimental, and descriptive vs. analytical, emphasise how the order was discovered (Çaparlar & Dönmez, 2016).

Arguing that the aspect of societal use or impact has been absent in the traditional classification criteria of scientific research, Zhang et al. (2021) included all areas of research and analysed the main aims of the research as reported in scientific publications by using another distinction. They defined *scientific progress* and *societal progress* as the two main alternatives, with the following criteria according to the presentation of scientific research in the publication:

- **Scientific progress**: *Statements of the aims and implications of the research refer to the advancement of knowledge in relation to previous research and/or potential new knowledge. External use of knowledge is not mentioned.*
- **Societal progress**: *Statements of the aims and implications of the research refer explicitly to external usefulness. The aim of contributing to scientific progress may also be stated, but not necessarily.*

Intriguingly, while Boyack et al. (2014) successfully applied machine learning models to categorise large-scale research articles into a four-level classification in the spectrum of basic vs. applied research, Zhang et al. (2021) examined the same models on their new proposed criteria but received negative outcomes. Clearly, it has been widely approved that embedding techniques could achieve much more prior performance in knowledge representation and its broad downstream tasks (e.g., topic extraction) in scientometrics (Wu et al., 2021; Zhang et al., 2018), and the rapid development of language models, including both pre-trained language models (e.g., BERT and its variants) and large language models (LLMs; e.g., GPT-3.5 and 4), has incredibly enhanced such capabilities (Achiam et al., 2023; Devlin et al., 2018). This motivated us to leverage the most recent advancements in language models to classify scientific research towards their research aims in scientific progress versus societal progress.

In this study, we introduced an automated method of research aim classification and conducted a comprehensive assessment of popular text classification techniques in scientometrics, including traditional text mining methods, pre-trained language models, and LLMs. Our experiments revealed that the fine-tuned Ssci-BERT-e4 model outperforms other approaches, including the fine-tuned GPT-curie model and instructed GPT-4. Additionally, we applied the fine-tuned model to analyse 2.3 million journal articles published in 2022, providing insights into the overall distribution of research aim classifications across different research categories, i.e., the Web of Science Subject Categories.

2. Data

This study utilised two datasets: A 868-record dataset (Dataset 1) for model training and examination, and a 2.3-million-record dataset (Dataset 2) for a case study.

- **Dataset 1**: This is a labelled dataset of 868 research records. In a previous study, some of our team members annotated the research aim of each paper as "scientific progress" or 'societal progress" by manually reading their titles and abstracts independently. Among them, 366 are labelled as "scientific" while the other 502 records are "societal" (Zhang et al., 2021).
- **Dataset 2**: It comprises all the 2.3 million research articles published in journals indexed by the Journal Citation Report (JCR) list in 2022. We retrieved them from the OpenAlex database (Priem et al., 2022).

The overall research framework is illustrated in Figure 1.

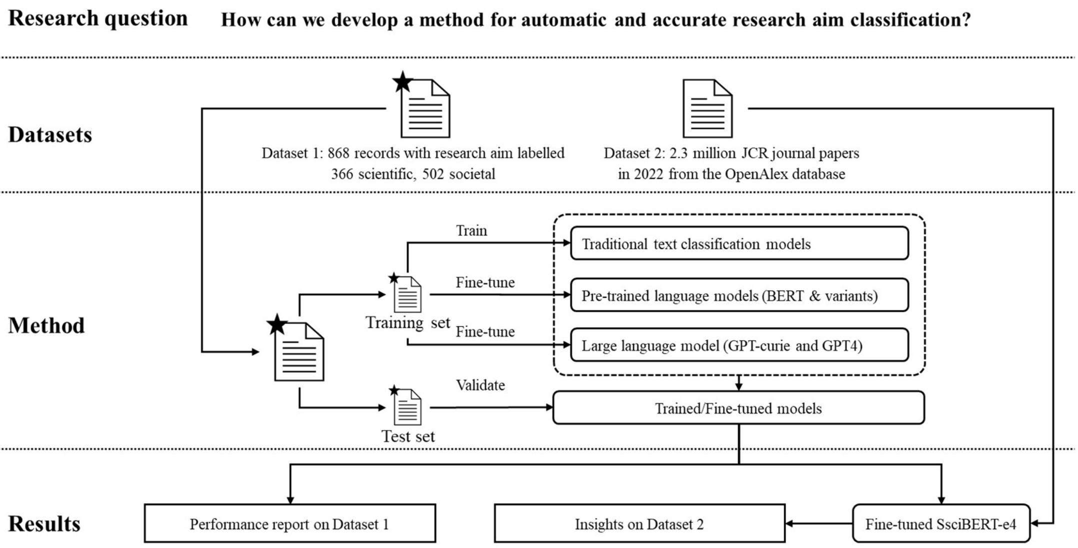
Research question
How can we develop a method for automatic and accurate research aim classification?
Datasets
Dataset 1: 868 records with research aim labelled
366 scientific, 502 societal
Dataset 2: 2.3 million JCR journal papers
in 2022 from the OpenAlex database
Method
Train
Fine-tune
Fine-tune
Training set
Traditional text classification models
Pre-trained language models (BERT & variants)
Large language model (GPT-curie and GPT4)
Validate
Test set
Trained/Fine-tuned models
Results
Performance report on Dataset 1
Insights on Dataset 2
Fine-tuned SsciBERT-e4

Figure 1. Research framework of research aim classification

3. Experimental design and results

3.1. Methods

We identified three prevalent types of text classification approaches and selected their representative models:

Traditional text mining models:
We followed the text mining models utilised in the classification study conducted by Boyack et al. (2014). These models leverage statistical patterns within text and represent text as a weighted aggregation of words or phrases, with weights typically determined by the word frequency or the Term Frequency-Inverse Document Frequency (TF-IDF) approach. While these methods are straightforward and thus efficient, they usually lack the ability to convey contextual information about the selected words or phrases. For this type, we selected word frequency (WF) and TF-IDF values as text features and combined them with three classifiers: multinomial Naïve Bayes (MNB), logistic regression (LR), and support vector classification (SVC) to establish baseline models. Then, we obtained six combinative approaches: WF-MNB, TF-IDF-MNB, WF-LR, TF-IDF-LR, WF-SVC, and TF-IDF-SVC.

Pre-trained language models
Pre-trained language models have been trained on extensive corpora, using specific tasks to acquire text representation or vectorization. These models utilise large-scale corpora to comprehend text semantics across various contexts and can be fine-tuned further by introducing domain-specific data and corpus, tailoring them to specific requirements. Unlike traditional methods, these models analyse word sequences to grasp contextual information, rendering them more adaptable and precise across diverse downstream tasks, e.g., text classification. Pioneering efforts in this field include the pre-trained Word2Vec (Mikolov et al., 2013) and subsequent advancements like the Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018).

Given the superiority of BERT models on fine-tuned tasks, we selected the original BERT and five variant models as candidates including BERT, SciBERT (Beltagy et al., 2019), BioBERT (Lee et al., 2020), Sentence-BERT (Reimers & Gurevych, 2019), SsciBERT, and Ssci-

SciBERT (Shen et al., 2023). The variant models of SciBERT, BioBET and Ssci-BERT were respectively pre-trained on scientific, biomedical, and social science corpora. The e2 and e4 mentioned in Ssci-BERT and Ssci-SciBERT refer to the times of related training epochs. Sentence-BERT is a BERT variant tailored for long-text and sentence classifications.

Large language models

LLMs are essentially one type of pre-trained language models but are separately listed due to their distinctly enormous size of parameters (billion level) and training corpora. Since the release of OpenAI GPT3.5/4 models and the chatbot ChatGPT, LLMs have demonstrated remarkable capabilities of human language comprehension, generation, and question answering. Such LLMs also hold the potential for text representation and classification (Achiam et al., 2023). In this work, we selected the GPT-curie model[1] for fine-tuning and the representative GPT-4 model for instruction-based classification.

3.2. Experiment setting

For data pre-processing, we initially concatenated the title and abstract of each record and used it as the training input. For the traditional text mining models, a streamlined process of lowercasing, stop words removal, stemming and lemmatisation was applied using the NLTK[2] Python package. For pre-trained and large language models, the original texts are directly given.

For the MNB, LR and SVC classifiers, we implemented them using the scikit-learn[3] Python package with default parameters. All BERT and variant models were accessed via the Huggingface[4] platform and fine-tuned on a server with three Quadro RTX 8000 48GB GPUs. The GPT-curie model was fine-tuned via OpenAI API[5]. For curating a high-quality prompt for instructing GPT 4, we fed our classification definitions to ChatGPT and let it generate the following prompt that it could understand:

> *You're an experienced annotator responsible for labelling research papers based on their titles and abstracts. If the aims and implications of the research focus on advancing knowledge relative to prior studies or potentially introducing new knowledge, the paper will be labelled as "scientific." Conversely, if the aims and implications explicitly emphasize external usefulness, it will be labelled as "societal." Just indicate "scientific" or "societal" for each instance.*

Dataset 1 was randomly divided into a training set and a test set according to the ratio of 8:2. Following the division, we fed the training data (along with labels) into candidate models and utilised the test data (no labels given as input) to validate the classification accuracy. The traditional and BERT models were trained/fine-tuned to output probabilities for both categories: *scientific progress* vs. *societal progress*. Given the sum up of the collective probability is 1, we labelled a record as the category with a probability above 0.5. Following the data division, training and testing operations were run independently five times, and we reported the mean of each evaluation metric.

[1] https://platform.openai.com/docs/models/gpt-base
[2] https://www.nltk.org/
[3] https://scikit-learn.org/
[4] https://huggingface.co/
[5] https://openai.com/pricing

3.3. Evaluation metrics

To measure the effectiveness of models on this classification task, we utilised the following four metrics:

- **Accuracy (A):** As the most fundamental and straightforward metric, it is computed as the ratio of correctly classified records to all records in the dataset.
- **Precision (P):** For a category, precision is defined as the ratio of true positive records to all records predicted as positive for that category.
- **Recall (R):** For a category, recall represents the ratio of true positive records to all records that belong to that category.
- **F1-score**: F1-score is the harmonic mean of precision and recall of a category. Additionally, macro F1 (F1-m) is the unweighted average of F1-scores across both categories, while weighted F1 (F1-w) is the F1-score averaged by the number of records in each category.

Note that we reported P, R, and F1 in a category bias (i.e., scientific vs. societal) to fully present the preference of the candidate models.

3.4. Experiment results

The experiment results on Dataset 1 are given in Table 1. The best results are highlighted in red, and the second-best results are underlined.

Table 1. Binary classification results on Dataset 1

	A	**Scientific**			**Societal**			**F1-m**	**F1-w**
		P	**R**	**F1**	**P**	**R**	**F1**		
TF-IDF-MNB	0.738	**0.968**	0.391	0.553	0.692	**0.991**	0.814	0.683	0.704
TF-IDF-LR	0.803	**0.877**	0.622	0.726	0.773	**0.937**	0.847	0.786	0.796
TF-IDF-SVC	0.814	0.789	0.763	0.774	0.833	0.852	0.841	0.808	0.813
WF-MNB	0.811	0.793	0.745	0.768	0.824	0.859	0.841	0.804	0.81
WF-LR	0.816	0.773	0.799	0.784	0.851	0.83	0.84	0.812	0.816
WF-SVC	0.809	0.765	0.794	0.776	0.847	0.823	0.833	0.805	0.809
F. BERT*	0.853	0.831	0.824	0.824	0.874	0.877	0.873	0.849	0.853
F. SciBERT	0.859	0.842	0.824	0.83	0.876	0.882	0.877	0.854	0.858
F. BioBERT	**0.869**	0.864	0.817	**0.839**	0.874	0.907	**0.889**	**0.864**	**0.868**
F. SentenceBERT	0.867	0.873	0.8	0.833	0.866	**0.915**	0.888	0.861	0.865
F. SS-B-e2	**0.882**	0.871	**0.848**	**0.858**	**0.894**	0.903	**0.897**	**0.878**	**0.881**
F. SS-SciB-e2	0.855	0.83	0.826	0.827	0.877	0.874	0.875	0.85	0.855
F. SS-B-e4	**0.882**	0.868	**0.851**	**0.858**	**0.895**	0.902	**0.897**	**0.878**	**0.881**
F. SS-SciB-e4	0.86	0.859	0.804	0.827	0.867	0.897	0.88	0.854	0.859
GPT-curie	0.834	0.845	0.801	0.822	0.888	0.859	0.872	0.565	0.851
Instructed GPT 4	0.81	0.754	0.817	0.783	0.858	0.806	0.831	0.807	0.811

Note: * all BERT models are fine-tuned on the training set.

Table 1 highlights the superior performance of the fine-tuned Ssci-BERT-e4 model compared to other approaches in the binary classification task. Traditional methods utilising TF-IDF, or word frequency tend to select weighted terms as representative features but omit crucial contextual information vital for this classification task. The distinctions between the two categories of research aims are not solely represented in word distribution but also heavily depend on how they are articulated in the original text and whether the aim describes external use beyond the original research domain. For instance, two articles in the same research topic may exhibit similar word distributions, yet their research aims could differ based on one or two sentences. As a result, relying solely on traditional term-only linear models (e.g., TF-IDF and WF) can lead to misclassification of such records, thereby compromising overall accuracy.

In contrast, the BERT variations yield superior results across all evaluation metrics, highlighting their adeptness at capturing contextual information. When compared to SciBERT, BioBERT, and SentenceBERT, the Ssci-BERT models (e2, e4), pre-trained on an additional corpus of social science research publications, exhibit noticeable enhancements in classification performance, particularly on societal records. Intriguingly, research articles towards societal progress do not necessarily belong to social science disciplines, but such extra training feeds clearly add values. As for the slight performance disparities between the fine-tuned Ssci-SciBERT and Ssci-BERT models, we attribute them to the likelihood that key sentences describing research aims in records are more inclined to be articulated in the style typical of social science literature.

The fine-tuned GPT-curie model and instructed GPT-4 model present acceptable results but remain less competitive than the fine-tuned BERT variant models. We attribute this observation to two potential factors after further analysis:1) Encounter with the LLM hallucination issue in our GPT-curie model practice, wherein the model generates random samples unrelated to scientific or societal contexts. This issue resulted in misclassification of records and consequently degraded all measured metrics. 2) Despite the capabilities of generative LLMs in generating text responses based on given prompts, their base models still demonstrate limited capability for domain-specific text classification tasks compared to encoder-decoder methods.

4. Case Study

To further validate the generalisability of the fine-tuned model (i.e., Ssci-BERT-e4, the best performer in our Dataset 1-based experiments) and investigate the distribution of research aims at a larger scale, we utilised the titles and abstracts from Dataset 2 as input for the model and generated all inferences. Rather than directly assigning scientific or societal categories to each paper, we preserved the original output, consisting of two probability scores for the two categories. This approach allows us to reflect the distribution of research aim tendencies across various subject categories.

Figure 2 illustrates the overall distribution of the 2.3 million research articles across the two categories. The X-axis represents the probability that the records belong to societal progress, while the Y-axis indicates the cumulative count of records. At the midpoint of 0.5, which serves as the division between scientific and societal categories, we observed that 32.9% of the research articles published in 2022 are attributed to societal progress. This may indicate that most studies tend to contribute to in-domain progress rather than societal contributions, at least most of them have chosen to not explicitly claim such contributions in their titles and abstracts.

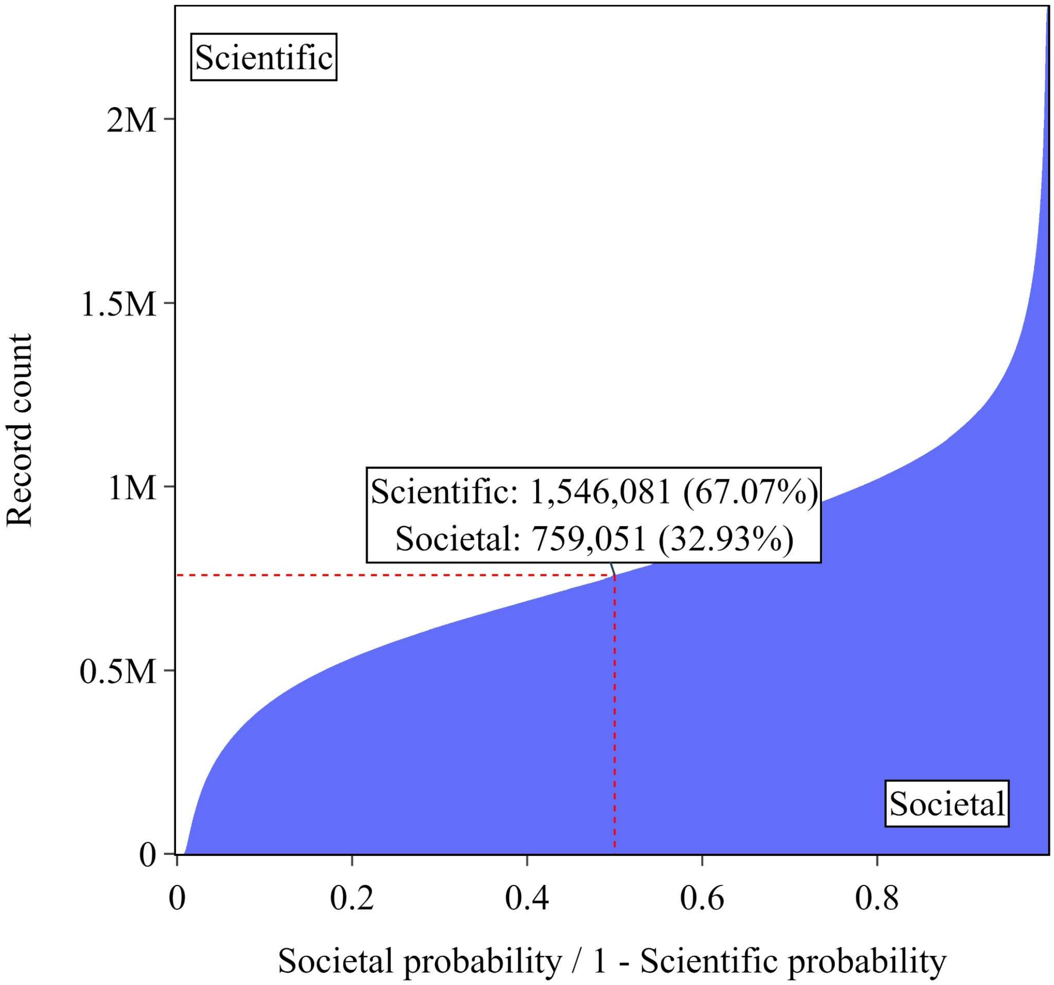

Figure 2. The distibution of scientific/societal probability on Dataset 2

Then we extracted results for the journal articles in the top fifteen Web of Science (WoS) subject categories with the largest number of publications in 2022 and organised the results in Figure 3. Not surprisingly, STEM-relevant subject categories generally demonstrate strong inclinations to scientific progress. Comparably, *Environmental sciences* and *Energy & Fuels* are two subject categories with more contributions towards societal progress. Clinical specialty-relevant subject categories, including *Oncology*, *Surgery*, *Cardiac & Cardiovascular Systems*, and *Clinical Neurology*, present above-average ratios to the societal category.

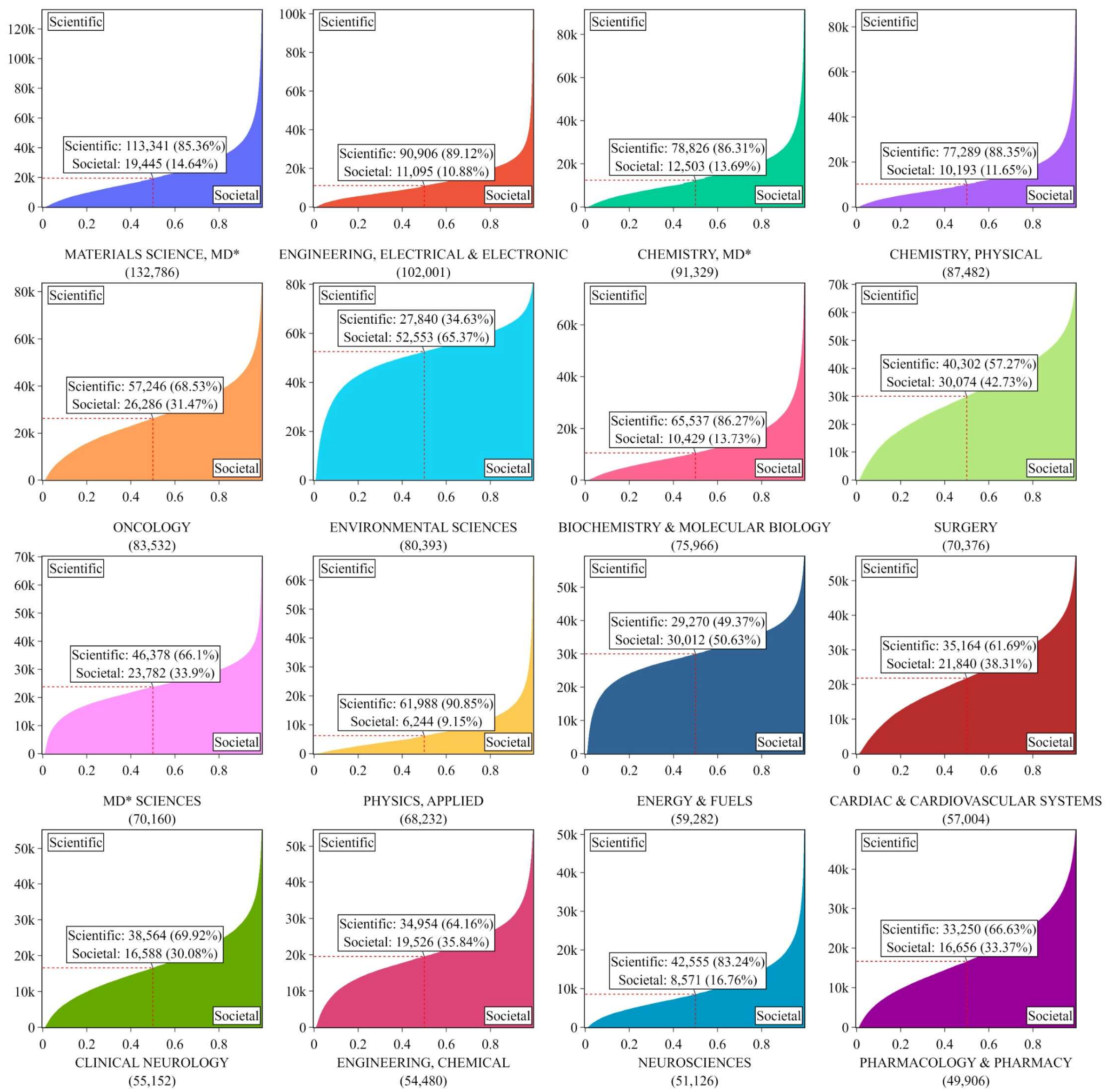

Figure 3. The distibution of scientific/societal probability for top 15 WoS subject categories

Note: *MD is the abbreviation for multidisciplinary. The number in the brackets under the name of a subject category indicates the number of publications in 2022 under that subject category.

Table 2 presents the top WoS subject categories with the largest proportion in scientific or societal progress. Despite observing some correlation between the scientific/societal classification and whether the subject category belongs to the Science Citation Index Expanded (SCIE) or the Social Science Citation Index (SSCI), we can still infer that the inclination to the societal category depends on how explicitly and directly the research of related subject categories can contribute to human society. For example, research in the SSCI subject categories of *Psychology, Mathematical*, *Psychology, Psychoanalysis*, and *Psychology, Experimental* (in the full list[6]), though belonging to the scope of social sciences, still are classified as scientific driven, since they seem to focus more on modelling development and innovation in psychology. Whilst subject categories like *Agricultural Economics & Policy*, *Substance Abuse* and *Nursing* are in the SCIE list, most of their research highlight contributions to external uses on social, environmental, and political matters.

[6] https://github.com/IntelligentBibliometrics/SSC/blob/main/class-category.xlsx

Table 2. Top 10 WoS subject categories with the largest ratio in two aim categories

Scientific Progress		
	# scientific*	# societal
Physics, Particles & Fields - SCIE	12,962	76
Mathematics, Applied – SCIE	28,554	223
Mathematics – SCIE	26,681	225
Logic – SCIE	1,014	10
Astronomy & Astrophysics - SCIE	22,278	266
Physics, Mathematical – SCIE	11,506	174
Crystallography – SCIE	4,980	98
Physics, Nuclear – SCIE	4,670	94
Psychology, Mathematical - SSCI	8,39	21
Societal Progress		
	# scientific	# societal
Green & Sustainable Science & Technology - SSCI	93	1,121
Agricultural Economics & Policy - SCIE	160	1,019
Family Studies - SSCI	652	3,775
Social Work - SSCI	634	3,652
Environmental Studies - SSCI	1,995	11,064
Public, Environmental & Occupational Health - SSCI	4,276	22,687
Hospitality, Leisure, Sport & Tourism - SSCI	710	3,668
Development Studies - SSCI	503	2,585
Substance Abuse - SSCI	765	3,851
Demography - SSCI	281	1,191

Note: # denotes the number of articles under the aim category.

5. Conclusions and discussion

In this study, we introduced an automatic classification approach for identifying research aims. We conducted a comparative analysis of mainstream text classification methods including traditional text mining models, pre-trained language models, and LLMs, and ultimately selected the fine-tuned Ssci-BERT-e4 model as our classification model for a case study on the 2.3 million journal articles published in 2022. Our experimental results revealed the following insights: 1) Traditional text mining models require minimal computational resources but exhibit compromised performance on complex text classification tasks that necessitate contextual information. 2) Fine-tuning domain-specific pre-trained BERT models remains a practical and cost-effective approach for text classification tasks compared to generative LLMs like GPT-4 and GPT base models, while the LLM's domain adaptability might be a key drawback in this classification task.

Certain limitations come with the current work. The scale of our fine-tuning dataset, comprising 868 records, is constrained by our limited human resources for annotation. The generalisability of the selected model has not been thoroughly examined. We anticipate three future directions for this work: 1) enhancing the scale of the training data by annotating more records or employing data synthesis strategies, 2) validating the model's generalisability on double-blinded annotated data from different domains, 3) investigating the correlation between research aim distribution and various bibliometric and societal indicators through large-scale analysis, including author genders, country incomes, etc.

Open science practices

The code and data of this paper is available upon request.

Author contributions

Mengjia Wu: Conceptualization, Formal Analysis, Methodology, Validation, Visualization, Writing – original draft.
Gunnar Sivertsen: Conceptualization, Data curation, Investigation, Validation, Writing – review & editing.
Lin Zhang: Conceptualization, Data curation, Investigation, Validation, Writing – review & editing.
Fan Qi: Conceptualization, Data curation, Investigation, Validation.
Yi Zhang: Conceptualization, Formal Analysis, Methodology, Writing – original draft, Writing – review & editing.

Competing interests

The authors declare that they have no competing interests.

Funding information

This work was supported by the Commonwealth Scientific and Industrial Research Organization (CSIRO), Australia, in conjunction with the National Science Foundation (NSF) of the United States, under CSIRO-NSF #2303037. The authors also acknowledge support from the National Natural Science Foundation of China (Grant No. 71974150; 72004169; 71573085), the Major Project of National Social Science Foundation of China (Grant No. 19ZDA349) and by the Research Council of Norway, Grant 256223, FORINNPOL.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., & Anadkat, S. (2023). GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Boyack, K. W., Patek, M., Ungar, L. H., Yoon, P., & Klavans, R. (2014). Classification of individual articles from all of science by research level. *Journal of Informetrics*, *8*(1), 1-12.

Çaparlar, C. Ö., & Dönmez, A. (2016). What is scientific research and how can it be done? *Turkish Journal of Anaesthesiology and Reanimation*, *44*(4), 212.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Hodes, R. (1974). Aims and methods of scientific research. In *For Dirk Struik: Scientific, Historical and Political Essays in Honor of Dirk J. Struik* (pp. 353-364). Springer.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, *36*(4), 1234-1240.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*.

Narin, F., Pinski, G., & Gee, H. H. (1976). Structure of the biomedical literature. *Journal of the American Society for Information Science*, *27*(1), 25-45.

Priem, J., Piwowar, H., & Orr, R. (2022). OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv preprint arXiv:2205.01833*.

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Shen, S., Liu, J., Lin, L., Huang, Y., Zhang, L., Liu, C., Feng, Y., & Wang, D. (2023). SsciBERT: A pre-trained language model for social science texts. *Scientometrics*, *128*(2), 1241-1263.

Wu, M., Zhang, Y., Zhang, G., & Lu, J. (2021). Exploring the genetic basis of diseases through a heterogeneous bibliometric network: A methodology and case study. *Technological Forecasting and Social Change*, *164*, 120513.

Zhang, L., Sivertsen, G., Du, H., Huang, Y., & Glänzel, W. (2021). Gender differences in the aims and impacts of research. *Scientometrics*, *126*, 8861-8886.

Zhang, Y., Lu, J., Liu, F., Liu, Q., Porter, A., Chen, H., & Zhang, G. (2018). Does deep learning help topic extraction? A kernel k-means clustering method with word embedding. *Journal of Informetrics*, *12*(4), 1099-1117.