Detecting Research Frontiers Based on Twitter (X)

Xiaomei Wang [1], Bor Jiang [2], Zhengyin Hu [3]

[1] Institutes of Science and Development, Chinese Academy of Sciences
[2] Beijing Jingwei Hirain Technologies Co., Inc.
[3] National Science Libray(Chengdu), Chinese Academy of Sciences

Nice, France, May 18-22, 2025

Xiaomei Wang

wangxm@casisd.cn

Professional Experience

- Researcher of Institutes of Science and Development, CAS
- Head of the Intelligence Analysis Technology and Data Platform team.
- Research primarily focuses on information analysis methods and technologies, scientific mapping analysis methods and tools, and Intelligent reasoning using scientific and technological big data.

Publications & Activities

- Mapping Science Strcture2022, China Science Press
- Mapping Technology Structure 2024, public release

Contents

1 Aims and contributions of our paper

In our paper, we aimed at:

1

- Design a method for identifying research frontiers based on user-generated content and in-depth behavioral interaction data from social media platform X.
- As a supplement to existing bibliometric methods, which are primarily paper-based, address issues such as lag in revealing emerging topics and insufficient dynamism, thus offering a complementary perspective.

Contributions of our study:

2

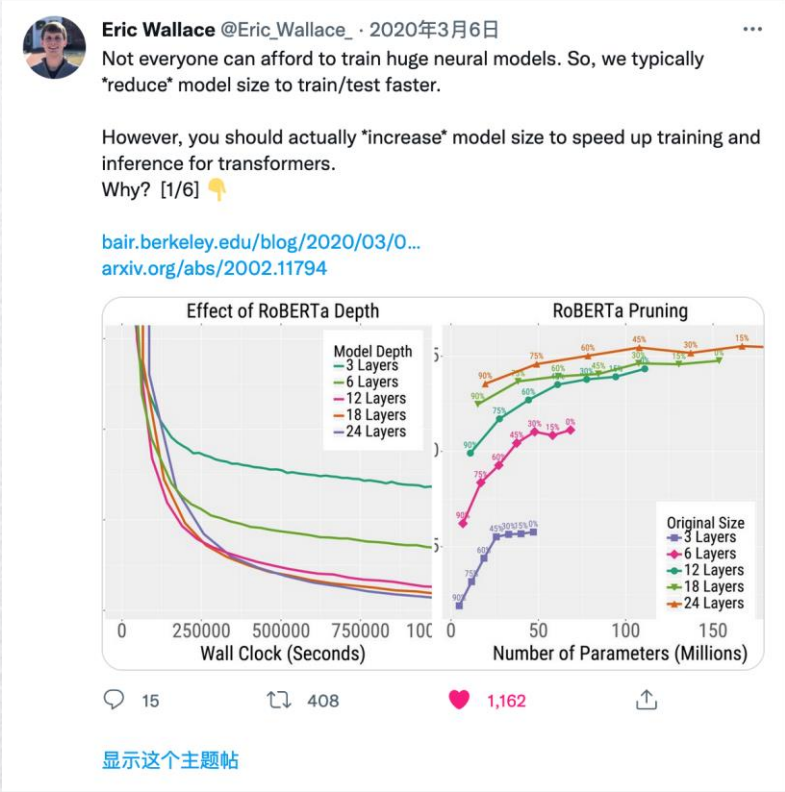
- Solve the issue of constructing domain-specific datasets on platform X.
- We propose a set of monitoring indicators for identifying research frontiers based on scholar influence and content impact, by analyzing the principles of research frontier identification in platform X, and conduct frontier detection in the field of NLP.

2.1 Social media is reshaping the model of academic communication

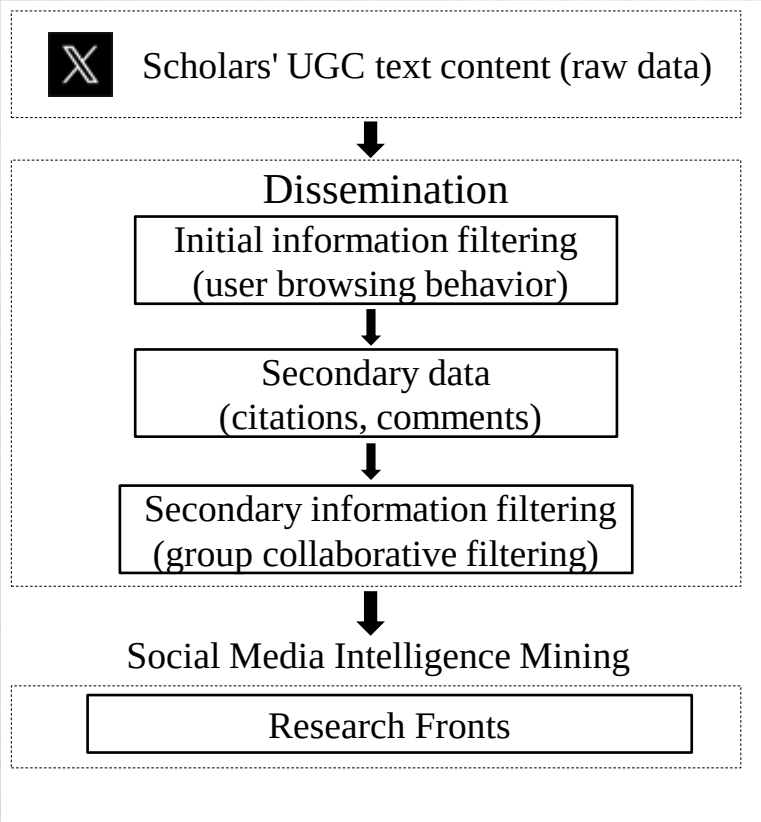
X has become an important platform for the academic community to discuss and analyze issues, share academic resources, and gain inspiration.

Scholars' behavior patterns on X:

- Disseminating research findings (promotion)
- Engaging in academic exchange
- Keeping up with front developments
- Expanding collaboration networks (social attribute)



2.2 Mechanism Analysis of Detecting Research Frontiers Using X



Sensitivity of social media to content

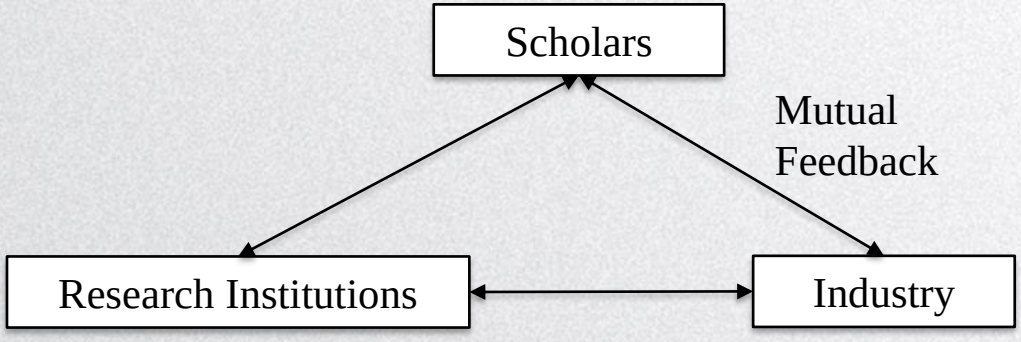
- High-value content, represented by novelty, innovation, and breakthrough research, is more likely to be disseminated and amplified by network effects.
- "Low-cost" social media interactions, such as likes and retweets, accelerate the dissemination and filtering efficiency of quality content.

2.2 Mechanism Analysis of Detecting Research Frontiers Using X

Group interaction behavior surrounding academic content provides a new way to monitor real-time dynamics in academic fields.

- **Consistent Interaction Behaviors:**
  - Groups with the similar value orientation and emotional attitudes express consistent interaction behaviors, even though factors such as cognitive differences and interests may lead to diverse interaction behaviors among individual users towards different content, at the level of complex social networks.
- **Collective Filtering and Evaluation:**
  - Communities with homogeneous knowledge structures and preferences can collectively filter out content with significant value contributions in their field on a systemic level. Each interaction behavior towards content in a specific academic field is considered a "vote" by the user for that content, thereby achieving the function of academic evaluation on social media.

2.3 Advantages of Identifying Research Frontiers Using X



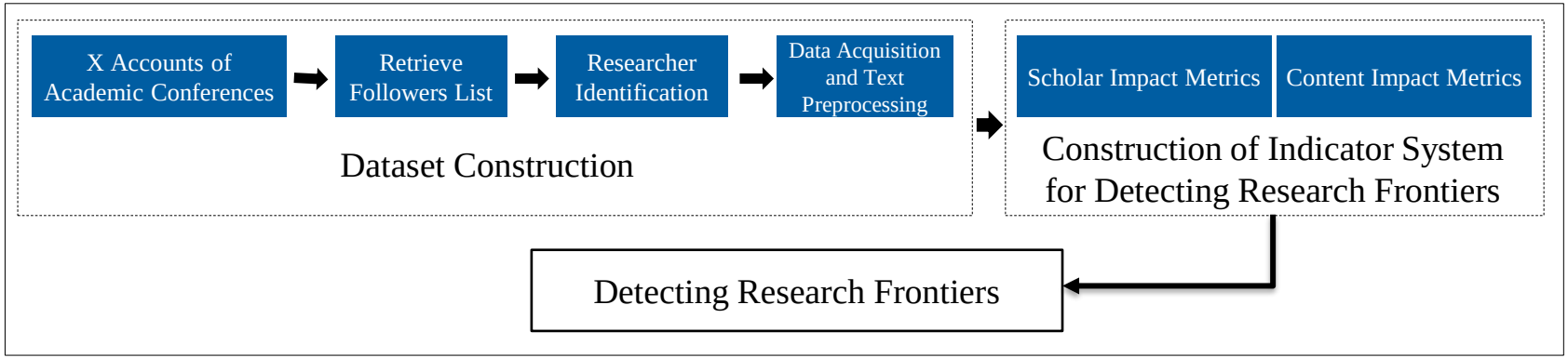
- Provide foundational computing power and training datasets support
- Model development
- Scalable application

X provides an information increment that traditional information sources cannot offer

- Compared to the strict requirements of formal publications, X, due to its open nature, provides a platform for scholars to discuss disciplinary issues or share immature research ideas or thoughts.
- X attracts industry users, broadening the scope of academic exchange and offering opportunities for industry-academia collaboration, communication, and the promotion of technological innovation.

The vast amount of topic discussion data on X around disciplinary fields, as well as the new developments released by institutions and industries, contains a wealth of information on the latest developments in various academic fields. This provides new and diverse perspectives for exploring and revealing research frontiers.

3.1 Methods for Detecting Research Frontiers Using X



Two Key Issues:

- How to construct a comprehensive dataset of disciplinary fields on the X platform.
- Construct an indicator system for detecting research frontiers based on the information organization and dissemination patterns of the X platform, specifically for X text data.
  - Although frontiers detection metrics based on scientific literature are mature, current indicator systems are only applicable to academic literature and citation networks, and are not suitable for social media environments.

3.2 Construction of X Domain Dataset

1. Challenges of Social Media's Openness

- **Diverse User Base**: X hosts a wide range of users; content quality varies greatly.
- **Research Difficulties**:
  - Hard to accurately collect discipline-specific data. Building comprehensive datasets is highly challenging.

2. Limitations of Traditional Keyword Search

- Low accuracy; recall and precision rates are unreliable.
- New terms emerge rapidly; keywords become outdated.

3. Our Solution

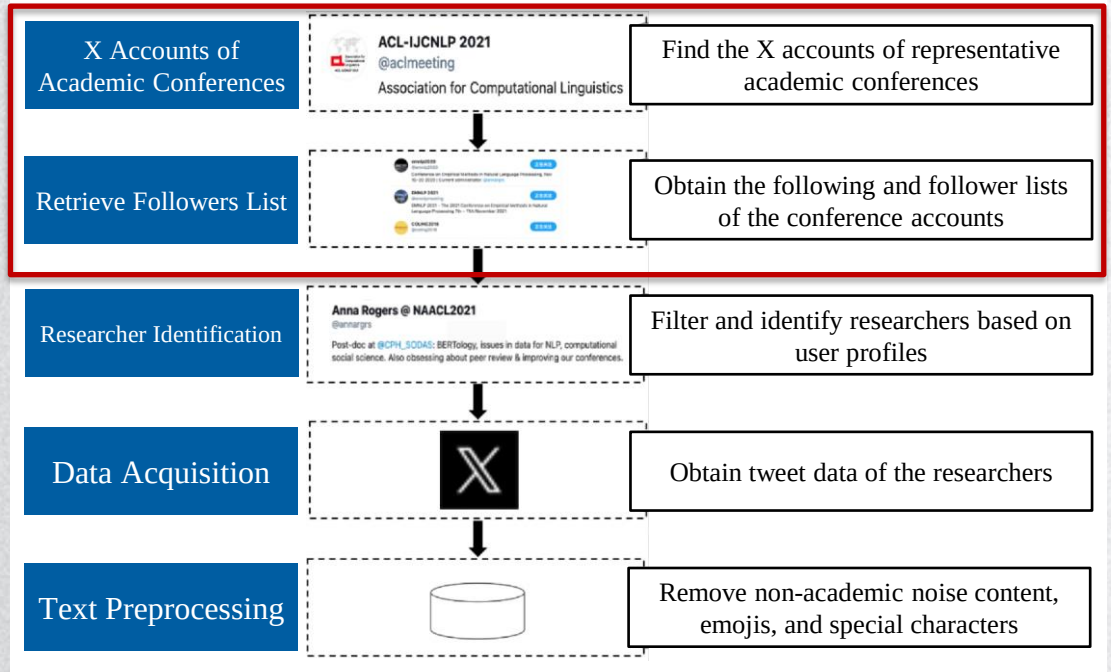
Focus on Relevant Entities:

- **Target Users**: Scholars, research institutions, related enterprises.
- **Data Collection**: Gather their tweets and interaction data.

Key Strategies:

- **User Identification**: Accurately locate relevant users on X.
- **Noise Filtering**: Eliminate content unrelated to the discipline.

3.2 Construction of X Domain Dataset



Identifying Scholarly Networks on X

Starting with Influential Conferences:

- Select highly influential academic conference accounts as "seed users".
- These accounts are central to the specific academic field.

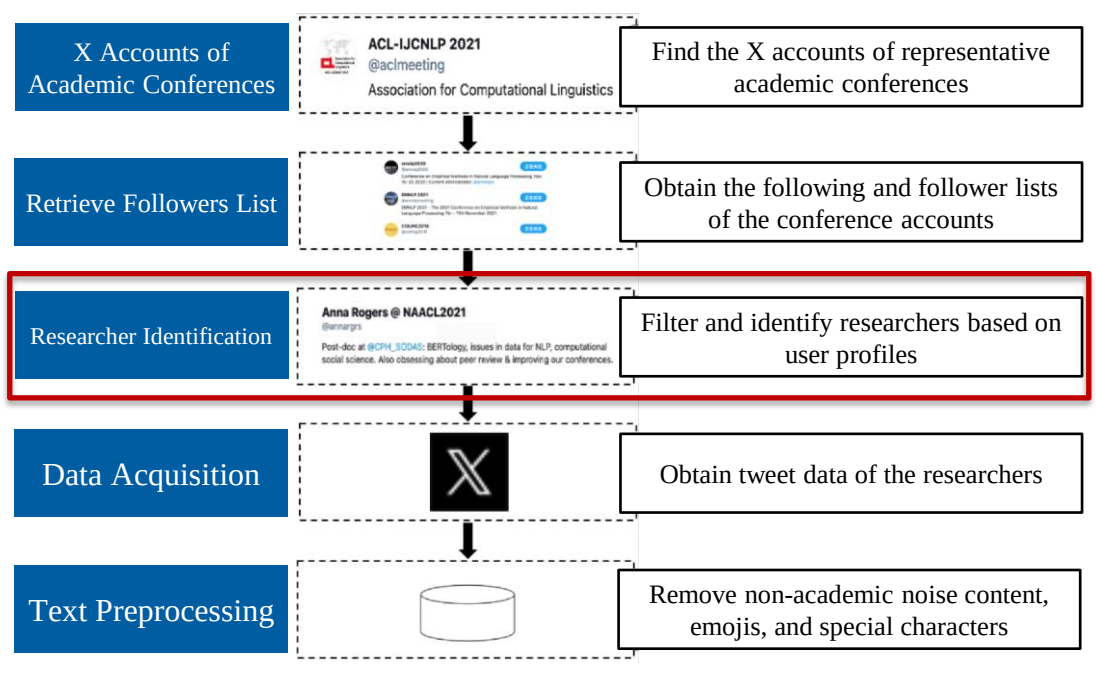
Interest-Based Following:

- Users tend to follow others who share their interests.
- Scholars following these conference accounts are likely involved in the same field.

Expanding the Social Network:

- Analyze the followers and following lists of the seed accounts.
- Identify and map the main group of scholars related to the academic field on X.

3.2 Construction of X Domain Dataset



Enhancing Data Reliability and Efficiency

•User for Profile Filtering:

- While most followers of academic conference X accounts are attendees or industry professionals, others may not be relevant.
- **Action**: Filter users based on their profiles to ensure data reliability and accuracy.

•Identifying Scholars via Profiles:

- Scholars typically use professional terms related to their occupation and field in their bios.
- Insight: Higher academic engagement on social media often means more research-related terms in profiles.

•Focusing on Highly Active Users:

- By targeting highly active users in a specific academic field, we capture high-value content filtered by the community.
- Benefit: Limiting data collection to these users improves acquisition efficiency.

3.3 Construction of Indicator System for Detecting Research Frontiers

Research shows that X interactions on academic content largely depend on the organization of the X content and the user attributes of its poster.

Primary Indicators	Secondary Indicators	Tertiary Indicators	Indicator Explanation
Content Influence	Value Influence	-	Number of Likes Received by a Tweet
	Dissemination influence	-	Number of Retweets and Quotes Received by a Tweet
	Novelty	-	Time Interval Between Tweet Posting and Current Time
Scholar Influence	Acceptance	audience attention	Total Number of Followers of the Scholar
		External Value Influence	Average Number of Likes on the Scholar's Original Tweets
		External Dissemination Influence	Average Number of Retweets on the Scholar's Original Tweets
	Activity Level	-	Average Time Interval Between the Scholar's Tweets
	Interactivity	Engagement	Number of Tweets Posted by the Scholar
		Depth of Engagement	Number of Tweets Commented on by the Scholar

Key Content Drivers

- **Core Determinant:**The ability to attract widespread attention and engagement hinges on significant value contribution.
- Other factors can only accelerate or decelerate the dissemination process.

Key Content Drivers

From the perspective of information consumers
From the perspective of information publishers.

- **Two Primary Indicators:**
  - **Content Influence Indicators**: Reflect value contribution driving academic communication and dissemination;
  - **Scholar Influence Indicators**: Derived from scholar elements.
- Sub-Indicators: Focus on the information organization patterns, dissemination patterns, and main behavior patterns of scholars on social media platform.

3.3 Construction of Indicator System for Detecting Research Frontiers

Weight Allocation Results

For the data samples, calculate each sample's score based on the quantitative indicator measurement methodology.

Primary Indicators	Weight	Secondary Indicators	Weight	Tertiary Indicators	Weight
Content Influence	0.6345	Value Influence	0.2605	-	-
		Dissemination influence	0.6333	-	-
		Novelty	0.1062	-	-
Scholar Influence	0.3655	Acceptance	0.6393	audience attention	0.1038
				External Value Influence	0.2311
				External Dissemination Influence	0.6651
		Activity Level	0.087	-	-
		Interactivity	0.2737	Engagement	0.5
				Depth of Engagement	0.5

Some quantitative description of the impact metric system

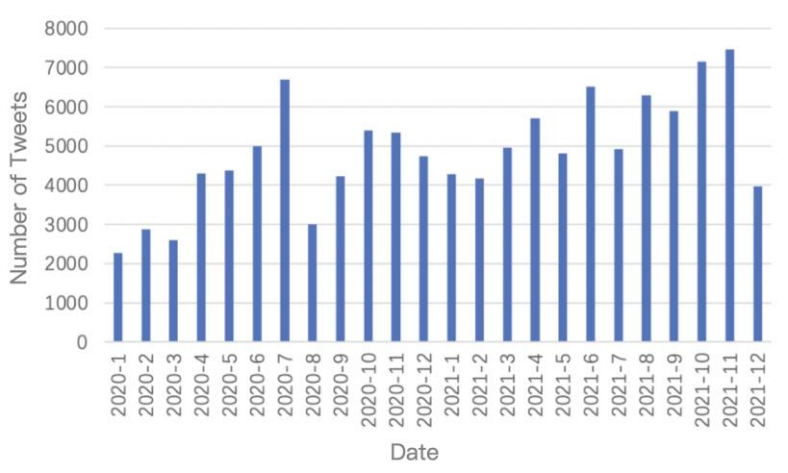
Novelty: $D_i = T_{now} - T_i$, D_i represents the temporal difference between the current time and the publication time of tweet i, where T_{now} denotes the current timestamp, T_i indicates the publication time of the i-th tweet, all timestamps are measured with daily precision.

External Value Influence: $L_u = \sum_{i=1}^{n} l_i / n$ represents the average like count of user u's original content, where n denotes the total number of user u's original tweets, l_i indicates the number of likes received by the i-th tweet.

External Dissemination Influence: $T_u = \sum_{i=1}^{n} t_i / n$ represents the average retweet count of user u's original content, where n denotes the total number of user u's original tweets, t_i indicates the number of retweets received by the i-th tweet.

Activity Level: $F_u = \sum_{i=1}^{n} (t_{i+1} - t_i) / n$ represents the activity level of user u, where n denotes the total number of tweets posted by user u, t_i indicates the publication time of the i-th tweet, all timestamps are measured with daily precision.

4.1 Obtaining Data



Some of the terms:

Professor, NLProc, Researcher, PhD, natural language processing

Model, language, speech, nlp, learn, text, research, paper, multilingual, AI, fine tune, train, state of the art, transformer

- **Seed account**: ACL, NAACL, CoNLL, COLING, EMNLP
- **User Selection:**
  - Initial Users: 2,037
  - Filtered Users: Removed users with fewer than 50 tweets and less than 100 followers; retained 425 users
- **Data Collection:**
  - Time Span: 2020-2021
  - Language: English only
  - Total Tweets Collected: 213,137
- **Data Filtering:**
  - Hashtags and Terms: Used #NLProc and common NLP terms
  - Final Tweets: After removing non-academic noise and duplicates, retained 107,225 tweets

4.2 Detecting Results

	Interpretation	Poster	Non-paper format		Interpretation	Poster	Non-paper format
1	The powerful capabilities of GPT-3 in code generation and automation	Open AI Researcher	√	11	Announcing the new research focus of the laboratory: developing tools that support collaborative construction of large models.	Professor at the University of North Carolina	√
2	GitHub Copilot：an AI programming assistant powered by OpenAI Codex, which has strong coding capabilities	Open AI Researcher	√	12	Replacing automated evaluation methods (such as ROUGE, BLEU) with human preferences as training objectives, using human feedback as rewards for reinforcement learning, and achieving performance in text summarization tasks that surpasses human levels.	Open AI	
3	The AI field is undergoing consolidation, with models from different areas converging into the Transformer architecture and being reused for various problems.	Director of Artificial Intelligence at Tesla	√	13	Testing the applications of GPT-3 and GPT-Neo in programming.	Researcher at Hugging Face	
4	Wav2vec can autonomously learn speech recognition in multiple languages from audio alone, without any speech-to-text training data, marking a significant breakthrough in speech AI.	Chief Technology Officer of Meta		14	BioMed Explorer: Applications of NLP models in the biomedical field.	Researcher at Google AI (Research and Health)	√
5	TruthfulQA: Testing the performance of language models in answering open-ended questions.	Oxford University researcher		15	The Current State and Future of Large Language Models	Researcher at Stanford University	
6	Call for attention to the application of NLP models in multiple languages.	Google AI Researcher	√	16	NLP open-source library: provides corpus management and evaluation functions	Researcher at Hugging Face	√
7	Multimodal text and image neural networks CLIP & DALL-E, used for text-to-image generation.	Open AI		17	To: By combining prompts with multi-task learning, it outperforms GPT-3 in downstream multi-task zero-shot performance testing.	Big Science	
8	ZeRO and the DeepSpeed library enables training of 100-billion-parameter deep learning models and has advanced Turing Natural Language Generation	Microsoft Research	√	18	Large language models are still progressing, and their capabilities continue to improve.	Deep Mind	
9	Large language models can synthesize and execute computer programs, solve math problems, and iteratively refine code through dialogue with humans, solving 60% of programming problems and 81% of math problems respectively.	Google Brain Researcher		19	REALM: A new paradigm of language pre-training models that enhances pre-trained language models with a knowledge retriever.	Google AI	
10	GSLM is a groundbreaking language model that eliminates the need for text in training, using raw audio recordings to generate expressive speech and representing natural language features through self-supervised learning and discrete tokens.	Meta AI		20	Call for focusing on potential linguistic phenomena rather than solely on the improvement of algorithms and models.	Professor at Bar-Ilan University	√

4.3 Validating Results

ML and NLP Research Highlights of 2021

This post summarizes progress across multiple impactful areas in ML and NLP in 2021.

Frontiers in ML and NLP Research	Is it NLP?	Does this study recognize?	Corresponding to the identification result number of this study
Universal Models	√	√	3、4
Massive Multi-task Learning	√	√	8、15、17、18
Beyond the Transformer	√	√	
Prompting	√	√	17
Efficient Methods	√	-	
Benchmarking	√	√	5
Conditional Image Generation	√	√	7
ML for Science	-	-*	
Program Synthesis	√	√	2、9、13
Bias	√	-	
Retrieval Augmentation	√	√	19
Token-free Models	√	√	
Temporal Adaptation	√	-	
The Importance of Data	√	√	16
Meta-learning	-	-	

*The method proposed in this paper successfully identifies the research frontier, and the research frontier score ranks in the top 20. However, since it does not belong to the NLP field, it is excluded during the preprocessing stage.

https://www.ruder.io/ml-highlights-2021

Comparison with Sebastian Ruder's Annual NLP Summary

- Benchmarking with Expert Insights

  This study references the annual summary blog post by Sebastian Ruder, a leading expert in NLP and Google scientist.

- **Ruder's Findings (2021):**

  15 research frontiers in ML & NLP, with 13 related to NLP.

- **Alignment with This Study:**

  8 out of the top 20 emerging frontiers identified in this study match Ruder's list.

  Additionally, our approach identified emerging frontiers such as Transformer architecture alternatives (ranked 22) and Token-Free models (ranked 31).

4.4 Limitations

1. Although Platform X provides better coverage of academic content compared to other social media platforms, it still falls short in comprehensively covering all research areas within specific fields.

- Some theoretically intensive studies tend to circulate primarily within the academic community and fail to gain significant attention on social media.

2.The dataset construction method may lead to omissions of some data:

- Users who do not follow conference accounts or those who neither publish nor engage with content may be excluded. However, this limitation is partially mitigated by the network effects of social media platforms, as high-value content (like novel innovations and groundbreaking research) is more likely to be disseminated through platform algorithms, amplified by user interactions, and captured by highly active experts in the field.

3. Compared to specialized academic publications, general social media platforms have limited academic content and contain significant non-academic noise, necessitating manual intervention during the analysis phase to interpret the content.

4.5 Conclusion and future work

Conclusion:

- The social media-based research frontier identification method proposed in this paper is capable of timely and forward-looking detection of research frontiers in the field of NLP.
- It represents a novel approach to revealing the dynamic development of research domain frontiers.

Future work:

- The scope of application should be expanded to validate the accuracy and effectiveness of this method across multiple disciplines.
- Further research will focus on developing efficient methods for acquiring academic content from X while minimizing the impact of noisy data from social media on the results.
- Additionally, combining traditional quantitative methods (such as journal articles, conference papers, etc.) with intelligence research will enable more accurate identification of research frontiers.

Thank You