Report on the 17th Meeting of Forum for Information Retrieval Evaluation (FIRE)

Thomas Mandl
University of Hildesheim
Germany
mandl@uni-hildesheim.de

Koyel Ghosh
SRM, Kattankulathur
India

Sukomal Pal
IIT (BHU) Varanasi
India

Krishna Tewari
IIT (BHU) Varanasi
India

Abstract

The Forum for Information Retrieval Evaluation (FIRE) is developing language resources for information access mainly in languages of India. The FIRE workshop was held in Varanasi in December 2025. FIRE included a scientific program and 12 evaluation tracks. These tracks provide recent insights into the domain of natural language processing and information retrieval. Many tracks are related to cross-lingual information access and analysis of social media data. More and more tasks are directed towards visual content. Languages of India are in the focus and in this edition, the following languages were addressed within the evaluation tracks: Hindi, Bengali, Gujarati, Bodo, Telugu, Tamil, Tulu, Kannada and Malayalam. Several tasks were offered for English.

Date: 17–20 December 2025.

Website: https://fire.irsi.org.in.

1 Overview

Systems for information access require corpora to train and evaluate them for each user and content language. The Forum for Information Retrieval Evaluation (FIRE) is focused on languages of the Indian subcontinent and was initiated in 2008. It complements large initiatives like TREC, CLEF, and NTCIR. FIRE has since evolved continuously to support and encourage research within the research communities. The FIRE conference was conducted from 17th to 20th December 2025 at the Indian Institute of Technology (BHU) Varanasi in India. The 17th edition of FIRE included 12 experimental tracks (shared tasks), a scientific conference program consisting of 14 papers, a PhD clinic, an industry session and 8 invited keynotes. For the conference track, a record number of submissions was received and the acceptance rate was around 30%. Proceedings appeared in the ACM digital library [Gangopadhyay et al., 2025] [1].

[1] https://dl.acm.org/conference/fire

The proceedings include papers accepted through peer review in the conference track and overview papers for each track. More than 80 papers will be included in the Working Notes proceedings consisting of the descriptions of the tracks and experiments carried out by the participants. Like previous editions, the Working Notes will be submitted to CEUR [Ghosh et al., 2025]. These participant papers describe experiments and show the performance of systems on the data sets offered in the tracks.

The shared tasks aim to facilitate research in IR and NLP. Several tracks were designated to NLP and IR tasks (Dravidian CodeMix, WILD, SqCLIR). Other tracks focused on access to scientific and technical information (SCi-High, CLMIR, IRSE). Content in the tourism domain is addressed by VATIKA. Further tracks provide specific corpora from social media platforms (CryptOQA, HASOC-meme, CMIR, PROMID task 3). Due to their accessibility, social media platforms are still a popular source for data. Problematic online content was also tackled by some tracks (PROMID, HASOC-meme, DravidianCodeMix). Two tracks were dedicated to Visual content (MUSIA, HASOC-meme).

A short overview on the previous edition of FIRE from 2024 is given in an editorial for a special section on highlight results of FIRE 2024 [Mandl and Majumder, 2026]. An overview of previous proceedings can be found at DBLP [2].

2 Conference Program

FIRE 2025 included the peer-reviewed conference track. This track aims to encourage novel research among early-stage researchers. The scientific program featured 14 selected presentations. Some of the papers explore topic related to India, ranging from language (indexing and searching for Sanskrit) to culture (access to Ayurveda knowledge). During the PhD clinic, 13 selected students had the opportunity to receive individual counseling from experienced researchers.

FIRE remains a in-person event with around 80 participants and only a few presentation were done online. This allowed frequent interaction of authors, students, track organizers and keynote speakers. The social program included a banquet which featured a performance of classical Indian music. The participants also had the opportunity to take a evening boat ride on the Ganges river to see ceremonies on the famous ghats. This part of the program was a unique experience in the spiritual city of Varanasi.

3 Keynote Talks

The program included 8 invites keynote talks.

The first one was delivered by **Joemen Jose** and had the title: "Beyond Hostility: Detecting Subtle and Overt Forms of Online Conflict with Multi-Objective Learning". It dealt with the toxic online environment and propagated conflict as a better concept for the domain of hate speech. The talk discussed the damage for brand communities which can occur due to conflict in social media channels. The analysis requires cooperation with marketing research.

Jaap Kamps talked about "How to Simplify Scientific Text (and Nothing More)?" This talk gave a report on a CLEF track on text simplification which is based on the Cochrane library of

[2] https://dblp.org/db/conf/fire/index.html

systematic reviews. Simple text is essential for inclusion of large groups of people who cannot process complex written text easily either because of the lack of expert knowledge or due to reading challenges. In the task, simple text summaries are aligned at the sentence level to the original abstract. Although models like BART and GPT obtained similar scores, it is remarkable that the work very different. It was observed that BART edited very conservatively and changes rather few words. On the other hand, GPT rephrased many parts, even in case it is not necessary. The talk proposed a RAG setup for text simplification.

Bhaskar Mitra gave a presentation titled: "Emancipatory Information Retrieval: Radically Reorienting Information Retrieval Research to Resist Corporate and Authoritarian Capture of our Information Ecosystems" The talk discussed the threats to democracy which big tech and AI pose. Big tech is empowered by access to tremendous amounts of data related to user behavior. The alignment of models to particular values is possible today and these can be set by large companies. The speaker mentioned venues for critical perspective on AI and mentioned the IR for Good track at ECIR and the Beyond workshop. He called for a emancipatory view on IR. The talk further elaborated the need to safeguard against surveillance, dehumanization of users and inequitable outcomes. Such an approach is necessary to mitigate human and environmental cost.

Utpal Garain spoke about "On Multi-role Alignment of Language Models". He discussed content moderation and the need for safe models. Blocking the creation of problematic output is required and strategies and evaluation results for blocking attacks were presented. The speaker elaborated on experiments which show that LLMs can be instructed to provide different answers to users for a medical data sets.

Ingo Frommholz talked about "Information Overload in Academia – Challenges and Opportunities for IR Research". He gave an overview about challenges related to scholarly information access. The issue of papermills and the detection of AI generated content (e.g. by finding tortured phrases) was discussed with many empirical examples.

Mark Sanderson gave a talk on "LLM as an Evaluation Forum?". The talk sketched the current controversial discussion within the community on using LLMs for relevance assessments. There are diverging empirical results on LLMs as a judge. The replacement of crowd workers by LLMs can help, however, it is yet unclear whether the top systems can be well distinguished by automated relevance judgments. The issue needs to be seen in a larger context. Humans are also active in other phases of the evaluation of IR systems, in particular, during the generation of information needs. Challenges for evaluation arise due to the potential of RAG approaches.

The keynote by **Liana Ermakova** is titled "Smart Tools, Critical Minds: How Users Shape the Reliability of AI". She reported on experiments on user Behavior with LLMs. One experiment showed how diverse user behavior is when it comes to the behavior regarding sources presented by an LLM. Furthermore, cognitive bias in search was discussed. LLMs seem to be subject to confirmation bias induced by assumptions given in a prompt. Furthermore, the talk showed also a framing and a position bias in selection tasks.

Arun Verma presented a keynote during the industry session titled: "NLP in Finance: Sentiment analysis, Topic Modeling and Geopolitical biases in LLMs". The talk reported on experiences at Bloomberg with the practical use of NLP tools. He gave the example of how annotations of persons and companies over long periods can create value added for industry level NLP analysis. Arun Verma showed the challenges or sentiment analysis in real-world scenarios, A main issues is avoiding noise. In production systems at Bloomberg manually curated hierarchical topic model

of products are used. Also for companies, a curated model is developed to provide sentiment analysis for investment decisions on the company level. Such sentiment analysis in large scale can currently only be done by a small model due to the need to efficient processing. Furthermore, also an in-house GPT at Bloomberg was built with 600 billion token training data.

4 Evaluation Tracks

The tracks form the core of FIRE. They are organized as shared tasks around a current topic of information access. In 2025, FIRE included 12 tracks which are described in the following subsections. Most tracks use standard platforms like Codabench or Kaggle for running the experiments. The most popular tracks had almost 20 participating groups. Further details on each track con be found in the respective overview paper in the Working Notes or in the abstract in the ACM proceedings.

4.1 Spoken-Query Cross-Lingual Information Retrieval for the Indic Languages (SqCLIR)

The track contributes to the core of cross-lingual information access. The vision is the development of systems which allow spoken input in Indian language and access to information in other languages. The data is a translation of the MS MARCO collection. SqCLIR includes short texts in five Indian languages: Hindi, Gujarati, Bengali, Kannada, and English. The task encompasses monolingual and cross-lingual retrieval settings across language pairs. It considers spoken queries with male and female voices. The primary objective is to assess retrieval effectiveness in a multilingual low-resource setting that reflects real-world language diversity.

4.2 Code-Mixed Information Retrieval from Social Media Data (CMIR)

Supriya Chanda presented the second edition of a track on code-mixing. In this particular track, Roman and Bengali script were mixed in social media posts. Often Bengali text was written in Roman script although there are no standard rules for transcription. Code-mixed communication is a widespread phenomenon worldwide. In India, code and language mixing are particularly common in social media platforms and systems need to address this phenomenon. The CMIR track provided a baseline system which already addressed some of the challenges of code-mixed text. 50 queries were given and 20 of them are test queries. These queries need to be run against a collection of more than 100.000 documents in order to retrieve relevant posts. overall, the performance was low. Successful approaches combined classic lexical methods like TFIDF and BM25 ranking with dense retrieval approaches. Rank fusion methods for combining diverse methods performed well. Over aggressive preprocessing can cause decrease in performance

4.3 Word-Level Identification of Languages in Dravidian Languages (WILD)

Also the track WILD faced Code mixed text as a reality in social media. It presented a data collection of 5 Dravidian languages. The track organizers pointed out that numbers and named entities are frequently subject to code-switching. Experiments and papers were submitted by 10 teams. They used diverse approaches to solve this task. Some relied on linguistic features like word length and capitalization. In addition, they used traditional approaches like TFIDF weights and applied classifiers like SVM. Others applied transformer based models to include context. LLMs are not used for this problem. The F1 scores are very different between languages. Future plans include an expansion of the task to the phrase level

4.4 Information Retrieval in Software Engineering (IRSE)

The automatic analysis and retrieval of code is an important task in Software Engineering. Previous research often focused on comment attributes, automated comment detection, and machine learning-based generation. IRSE intends to create reliable benchmarks for evaluating system in this domain. Task 1 is dedicated to Software Metadata Classification using Generative AI. A binary code comment quality classification model needs to be augmented with generated code and comment pairs that can improve the accuracy of the model. The data consists of 9048 pairs of code and comments written in C, labeled as either Useful or Not Useful. Task 2 is titled "Bring your own LLM". Here, the organizers provide a parent LLM and a RAG architecture designed to generate metadata for C source code. The RAG system includes question-answer pairs and comments focused on understanding and interpreting C code snippets. Participants had to scrape C code and associated comments from codebases and use this data to build a small language model from scratch. The comments generated for a set of 50 C programs were compared to human-written reference comments for the same code.

4.5 Cross-Lingual Mathematical Information Retrieval (CLMIR)

Cross-Lingual Mathematical Information Retrieval (CLMIR) aims retrieve mathematical content across different languages. CLMIR focuses on English-Hindi cross-lingual retrieval, helping users to find mathematical content across languages. The challenge in building such a system lies in accurately translating mathematical expressions and text. The dataset for the CLMIR 2025 task is curated from the Math Stack Exchange corpus from ARQMath-1 and contains approximately 39,862 instances (Training Data). The dataset is formatted to include the body of scientific information which contains mathematical equations, expressions, and supporting textual descriptions in Hindi and its associated search ID. To assess the performance of participants, 10 formula and text-based queries in English language (Validation Data) and 50 formula and text-based queries in English (Test Data) were provided.

4.6 Hate Speech and Offensive Content Identification in Memes in Bengali, Hindi, Gujarati and Bodo (HASOC-meme)

The task of hate speech detection increasingly necessitates the analysis of multimodal data, as harmful online content often exploits the combination of text and images to convey hateful messages in subtle or coded forms. In many instances, the textual content alone may appear not problematic. However, when two modalities come together, a hateful message may emerge. Analyzing both textual and visual content concurrently enables a more comprehensive understanding of the context in which hate speech is embedded. For this track, memes in Bangali, Hindi, Gujarati and Bodo were collected form social media sites. A 4-level annotation scheme was used for the analysis: Sentiment, sarcasm, vulgarity and abuse. Some memes were included that did not contain text. Participants used transformer based systems and Vision-language models to analyze both modalities jointly. Mostly, models including Indian languages like IndicBERT and XLM were used for embedding the text. For the visual information, vision transformer and ResNet architectures were employed.

4.7 Misinformation Detection and Prompt Recovery (PROMID)

The detection of misinformation is facing further challenges as LLMs are being used pervasively. PROMID is focusing on issues related to understanding how misinformation can be detected at diverse stages. The first task aims at finding factual incorrectness in machine-generated cross-lingual summaries. The second task in PROMID is dedicated to the classification of prompts which are likely to produce hallucinated content. Task 3 required the identification of misinformation classification for the 2022 Russo-Ukrainian conflict at the tweet-level. RoBERTa was a popular model among participants.

4.8 Offensive Language Identification in Dravidian Languages (DravidianCodeMix)

Offensive language detection is a critical task in natural language processing, particularly in the context of online discourse, where harmful content can spread rapidly. Identifying offensive language is challenging due to the varied ways in which offense is conveyed. Detection quality for many low resource languages is still not satisfying. A binary hate speech detection task was provided in Tamil, Malayalam, Kannada and Tulu. The data consisted of 35.139 posts from social media platforms for Tamil and smaller amounts of data for the other languages. The challenge for the classification is complex due to code-mixing between languages, as well as the use of non-native scripts.

4.9 Opinion Extraction and Question Answering from CryptoCurrency-Related Tweets and Reddit posts (CryptOQA)

Kripabandhu Ghosh presented the second edition of a track on the analysis of posts Opinions on Crypto currencies. The data consisted of over 30.000 short texts and underwent annotation in 3 levels. The track included 2 sub-tasks: Opinion Classification from CryptoCurrency related

Social Media Posts and Question Answering from CryptoCurrency related Social Media Post. For the second sub-task, participants received a set of question-comment pairs and had to identify the relevance of the comment with respect to a given set of cryptocurrency-related social media posts. Successful systems used systems like BERT but also newer models like GEMMA. The best participant achieved a perfect performance with a F1 score of 1.0. For the model CRYPTOBERT, a lower performance than for BERT was measured.

4.10 Research Highlight Generation from Scientific Papers (SciHigh)

The track SciHigh is related to Document analysis and the goal is the creation of very short abstracts of scientific articles. Around 10.000 documents were available for training. The pairs include abstracts and highlights written by authors. From this data, 1985 pairs were reserved for validation and 1840 for testing. ROUGE is used as evaluation metric, Participants encountered length restrictions of models. Models like Pegasus, T5 and BART were used and summarization techniques were applied. Also, pre-training models on data from the PubMed database was done. An expansion of the track to Indian languages is envisioned.

4.11 Varanasi Tourism in Question Answer System (VATIKA)

This text-based track is situated in the tourism domain. VATIKA adds a local touch to the shared tasks, as the topic is reliable tourist information for Varanasi. The city of Varanasi is a highly popular tourist attraction especially for many national tourists. The organizers provided 7900 tourist contexts and accordingly 19.900 QA pairs in Hindi. The Information on sights in and around Varanasi was collected from relevant sites of sights and tourism portals. Field visits were carried out to verify the information. Especially, for smaller sights, online information was often not available or accurate. The data was categorized into 10 tourism subdomains (like temples, museums or agencies). Answers were checked by human experts. F1 scores for correct answers were used for evaluating the systems and BLEU scores were also reported. Overall, 10 participants submitted working notes papers. Several groups used data augmentation by LLMs. An expansion of the track to further Indian languages is planned.

4.12 Multilingual Story Illustration: Bridging Cultures through AI Artistry (MUSIA)

Anshita Malviya presented this track on children story illustration. The research question behind this track is how well AI methods can read a children short story and generate a sequence of images illustrating the story. The results are expected to capture important moments in the narrative. The organizers provided 39 stories in Hindi and 39 in English. A 3-level evaluation scheme was used after submission. It judged the consistency between panels (time, characters), the visual quality and the fit to the story (relevance). Diffusion models were very popular among participants. The cross panel consistency was much lower than the visual quality. Models were not culturally aware and created Western style illustrations even for stories which were set in other contexts. Parth Patel described how the created elaborate prompts. Gemini was applied to generate prompts for the diffusion model Imagen. These prompts include style, color and further

visual details. Other participants extracted and weighed entities in the stories to obtain items for prompting. Open models seemed to perform less well than large closed models. Other groups used abstractive summarization to capture story parts and use the outcome in prompts. Most groups had to automatically translate from Hindi to English in order to generate prompts for images because no good models are available for Indian languages. Cross-attention was applied to maintain the consistency of characters among several images. A hard challenge was posed by multi-character scenes.

5 Outlook

The next edition of FIRE in 2026 will follow roughly the following timeline:

- April: Track proposals are due
- May: Track acceptance notification, track websites are available and release of training data
- June: Test data release and run submission deadline
- July: Track results declaration
- August: Working notes due
- September: Camera-ready copies of working notes due
- December: FIRE Conference at Dhirubhai Ambani University (DAU) in Gandhinagar, India (formerly DAIICT) [3]

References

Surupendu Gangopadhyay, Debasis Ganguly, and Debarshi Kumar Sanyal, editors. *FIRE '25: Proceedings of the 17th annual meeting of the Forum for Information Retrieval Evaluation*, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400723247. doi: 10.1145/3777867. URL https://dl.acm.org/doi/proceedings/10.1145/3777867.

Kripabandhu Ghosh, Thomas Mandl, Prasenjit Majumder, and Debasis Ganguly, editors. *Working Notes of FIRE 2024 - Forum for Information Retrieval Evaluation (FIRE-WN 2024), Gandhinagar, India, December 12-15, 2024.*, volume 4054 of *CEUR Workshop Proceedings*, 2025. CEUR-WS.org. URL https://ceur-ws.org/Vol-4054/.

Thomas Mandl and Prasenjit Majumder. Editorial: Special Section Forum for Information Retrieval Evaluation (FIRE) 2024. *Pattern Recognition Letters*, 199:285–287, 2026. ISSN 0167-8655. doi: 10.1016/j.patrec.2025.10.012. URL https://doi.org/10.1016/j.patrec.2025.10.012.

[3] https://www.daiict.ac.in