*Journal of Harbin Engineering University*
*ISSN: 1006-7043*

Vol 47 No. 03
March 2026

# InsightFlow: An Intelligent System for Automated Knowledge Discovery and Targeted Information Sharing

**Ms. Rashmika S B, Dr. R. Sabitha,**

Department of M.Tech Computer science and engineering,

Sri Krishna College of Engineering and Technology, Kuniamuthur, Coimbatore, Tamil Nadu - 641008

professor, Department of Computer science and engineering,

Sri Krishna College of Engineering and Technology, Kuniamuthur, Coimbatore, Tamil Nadu - 641008

***Abstract*** — In the digital age, they are struggling to find appropriate and quality learning resources with this tremendous amount of online educational materials. Manual search, filtering, and sharing is time-consuming and ineffective, it can lead to information overload & low learning efficiency. Traditional keyword-based search systems cannot represent the semantic information of content, and the current content-aggregation platform does not provide a quality review mechanism and target-release system either, which have led to an issue that users in organizations cannot timely share their knowledge with peer groups completely.

To overcome these challenges, InsightFlow aims to establish an automatic system of intelligent knowledge discovery and dissemination. The system fetches articles from some educational websites and adds them to a structured database. The Informative Value of the Content With the help of NLP methods, its content is examined both for quality, and semantic relevance. Abstractive summarization with Transformer produces short, readable summaries that convey key concepts while minimizing cognitive overhead. The summaries are then matched with the peer groups depending on embedding-based similarity scores, which means that the domain-specific and personalized summary is delivered. All of this is orchestrated by automation tools so content ingestion, processing, and distribution for consumption are predominantly done with little human activity.

Experimental results show that InsightFlow could greatly enhance relevance of result, quality of clips transferred, summarization capability and targeted delivering effect instead of the traditional keyword-based approaches as well as rule based methods. The platform facilitates real-time sharing of knowledge with systems like Telegram, enhancing accessibility and collaboration in communities of learning. InsightFlow offers scalable, intelligent and completely automatic technology to link up enormous availability of online content at the scale of the web with effective use of educational knowledge.

***Keywords*** *- Knowledge Discovery, NLP, Abstractive Summarization, Semantic Relevance, Automated Dissemination, Peer Group Mapping, Educational Content.*

## I. INTRODUCTION

Online Source Materials of Education and Techniques: The Good, the Bad, And Public Policy The rapid development of online education resources and technical has brought about new opportunities and presented challenges to learners. Although there is a wealth of content available on websites, blogs and learning platforms, students often struggle to find the right kind of reliable and relevant content. The large number of resources can cause overload, redundancy, and misapplications with manual search, filtering and assessing. Traditional methods of content discovery (such as keyword search engines and rule-based aggregation systems) have typically struggled with capturing the semantic context around educational content leading to irrelevant or low-value responses. Furthermore, most available solutions do not offer automated facilities for the summarization or targeted dissemination of

knowledge leading to a very time-consuming and non-uniform way for knowledge sharing among peer members.

Given the above limitations, there is a lack of intelligent and automated system that can continuously habitually discover, evaluate, summarize, and disseminate educational contents in an organized way as well as context-adapted. A systems of this kind should guarantee the correctness, brevity and specialization of the knowledge it provides to learners without requiring extensive manual work. It should facilitate peer learning by distributing content to the relevant peer groups as well, enhancing both learner engagement and learning effectiveness.

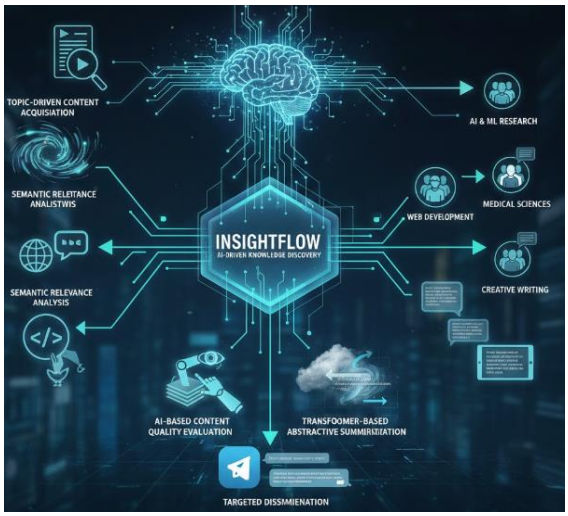


***Fig. 1. The InsightFlow System Overview***

In this paper we introduce InsightFlow, an AI-based system for automated knowledge discovery and information targeting. The platform combines a topic-driven content harvesting, semantic relevance scoring, AI-based relevance evaluation, and transformer-based abstractive summarization in an end-to-end workflow. Domain-specific peer groups are created by an embedding-based similarity approach and content is shared among them in conversational platforms like Telegram. Through the powerful synergy of semantic intelligence, workflow automation and targeted data dissemination, InsightFlow solves the problems of information glut and ineffective knowledge sharing in a way that scales to meet the needs of schools, training centers, and online learning groups.

## II. LITERATURE REVIEW

Recent studies have investigated approaches for semantic content retrieval and AI-based knowledge discovery mechanism to improve the access of educational information and personalization of content. The embedding-based semantic search models have been widely adopted and claimed as effective techniques to capture the contextual meanings in academia content and enhance the relevance of retrieved results [1], [2]. Studies by Sajja et al. [3] and Granata et al. [6] confirmed that hybrid retrieval-augmented generation approaches and semantic re-ranking methods could optimize the match of educational resources to learner needs. However, most of these works focus only on retrieval, and they usually do not incorporate automatic summary generation or propagate the generated links to peer groups.

Due to AI, there has been huge development in text summarization and semantic content handling in educational portals. The transformer-based models namely, BART [5], T5 and PEGASUS achieve commendable performance in generating short summaries from long educational articles. Liu et al. [11] and Wang et al. [13] underlined semantic embedding's for content-mapping towards personalized learner profiles, which guarantee both contextual coherence and domain relevance. But such efforts usually are standalone and do not feature end-to-end content automation from discovery, summarization to delivery.

Semantically linked knowledge graphs and ontology-driven systems improve content organization as well as targeted recommendations. Studies by Ramteja et al. [7] and Kalyani et al. [9] proved that structured semantic models allow personalized educational recommendations increasing learner engagement. Novel works emphasize the increasing relevance of AI-based semantic platforms in scalable knowledge management and automated education content dissemination [12] -[15]. In general, prior-art works focus on retrieval, summarization and recommendation independently, which indicates the necessity for an integrated system designed to integrate semantic-based relevance analysis and AI-powered inference summarization with peer-group

mapping and automatic knowledge delivery as proposed in this research.

## III. METHODOLOGICAL APPROACH

The methodology followed in this work aims to accommodate an organized and automated mechanism for facilitating educational information extraction/knowledge discovery and targeted information dissemination. In the methodology, it applies a sequential and modular process that augments each step in isolation and communicates them through an automatic execution. This architecture provides reliability, flexibility, with a minimum human intervention in the knowledge processing pipeline.

### *3.1 Overall Methodological Framework*

The overall methodological framework is comprised of seven linked stages: data source identification, content collection, pre-processing and normalization, semantic relevance assessment (SRA), content quality assessment (CQA), text summarization and knowledge classification with distribution. The pipeline starts with the collection of topic-guided contents from reliable educational resources and follows up several filtering and transformation levels in order to obtain a more focused, accurate and compact information. Such an automation system orchestrates the execution of all stages for the continual and scalable process applicable to a learning environment such as an academic system.


***Fig. 2. Methodological workflow of the proposed knowledge discovery and information dissemination framework***

### *3.2 Data Source Identification and Collection*

The first stage consists of predetermining domains and sources by the administrator. These sources span from reputable education websites and technical blogs, to learning web sites for a variety of topics. Scheduled crawlers collect articles and associated metadata in accordance with the specified sources. This process of recording data is also a way to control the type and quality of the content that is being ingested directly (untrusted or non-educational content).

### *3.3 Data Preprocessing and Normalization*

Web content collection is then preprocessed for useful textual information. The HTML tags, scripts, advertisements and navigational elements are removed to reduce the noise. The extracted text is later normalized processing the encoding, removing special characters and extra white space. Duplicated and near-duplicate articles are detected and excluded to prevent entering a cycle of knowledge propagation. The cleaned and organized text is put away in a database for analysis.

### *3.4 Semantic Relevance Assessment*

In order to provide the context relevance of each article, we conduct semantic relevance assessment via embedding-based similarity analysis. The contents of the articles and predefined topic descriptions are converted into vectors. The cosine similarity is calculated among these vectors to get M indicators of contextual alignment. Relevant articles are those which meet a predefined similarity threshold and are sent for downstream processing, while non-relevant ones are filtered out. This semantic comparison has the advantage of context-sensitive natural language meaning, 1 beyond what is possible with keyword-based term matching.

### *3.5 Content Quality Evaluation*

After assessing the relevance, there is also a stage which we apply and it's to evaluate the quality of the content in order to try to discard those articles with low-quality or non-informative articles. This appraisal includes, language coherence, structure completion, topic cilusivity and redundancy. Articles not passing the quality criteria are pruned from the pipeline. We make this step-up to prevent retaining of untrustworthy and educationally useful content which is essential for knowledge sharing in an academic environment.

### *3.6 Text Summarization Process*

The filtered articles are summarized using an abstractive summarization technique to obtain insightful and condensed summaries. Long articles are broken into manageable-length text and limited

to a maximum number of words allowed for the model input. The summary partials of each segment are summarized, and the following compressed summaries that are then combined to output the final summary. This procedure retains the essential semantic content while enormously reducing text length making it easier for learners to process the information.

*3.7 Knowledge Classification and Distribution*

In the last phase, summarized information is classified and aligned with relevant groups of peers according to domain-based classifications. To achieve accurate matching, semantic similarity calculation of the article summary vs. peer group profiles is carried out. After classification, the abstracts are disseminated with the corresponding subject groups to peers by communication services. By focusing the push transmissions, we make sure that only those files that are somehow relevant to a learning task are transferred just before they need to be learned.

**IV. ARCHITECTURAL IMPLEMENTATION**

The proposed system architecture consists of a modular and layered structure that can facilitate automated knowledge discovery, and efficient targeted information dissemination. Each architectural layer has its own duties and communicates with other layers through well-defined interfaces, to support scalability as well as easy maintenance.

The content harvesting layer obtains educational articles from predefined and trusted sources using spidering. The aggregated content is stored and managed in organized database via the data storage and management layer to provide efficient retrieval and control of duplicate information.

The core processing layer is responsible for performing semantic processing in which context relevance and content quality are used to filter the relevant educational material. All processing stages are scheduled on the automation and orchestration layer, ensuring hands-off operation with as little human intervention as possible.

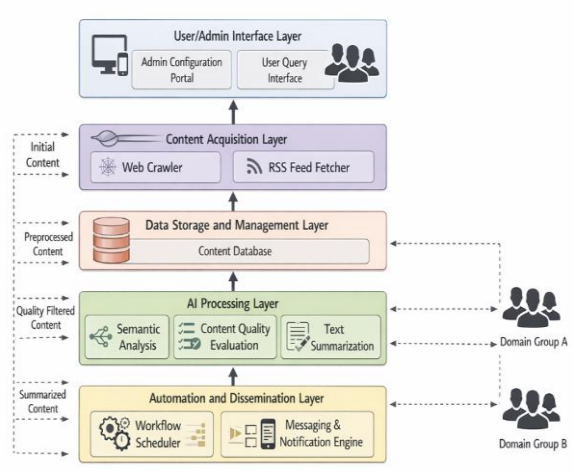


***Fig. 3. System architecture of the proposed educational content discovery and dissemination framework***

Finally, the high-end of knowledge diffusion provides aggregated and categorized content to relevant peer groups using communication services. Such a tiered architecture design provides reliable management as well as efficient processing and directed knowledge delivery appropriate for academic learning environments.

**V. ALGORITHMS AND MATHEMATICAL FORMULATION**

The presented framework offers several mathematical models and core algorithms for semantic relevance assessment to offer targeted knowledge sharing, content filtering, and summarization.

*Algorithm 1: Topic-Driven Content Acquisition*

**Input:**
Predefined topic set $T=\{t_1,t_2,\ldots,t_n\}$

$T=\{t_1,t_2,\ldots,t_n\}$, source URL list S.

**Output:**
Raw article set A

**Steps:**

- Initialize topic list T.
- Crawl educational content regularly, for each source $s \in S$.
- Extract raw article text and metadata.
- Save the articles to the database.

*Algorithm 2: Data Preprocessing and Normalization*

**Input:**
Raw article set A

**Output:**
Clean article set $A_c$

**Steps:**

- Removes HTML tags and scripts and advertising.
- Remove special character and extra spaces from string text.
- Find duplicate articles by comparing similarity threshold.
- Retain unique, clean articles.

*Algorithm 3: Semantic Relevance Assessment*

This algorithm evaluates whether a story is related on-context to a preset theme.

***Mathematical Formulation***

Let:

- $E_a$ = article embedding vector
- $E_t$= embedding vector of topic

Cosine similarity is computed as:

Sim(a,t) = (Ea · Et) / (||Ea|| ||Et||)

Decision rule:

If Sim(a,t) ≥ θ, article is relevant.

Where θ is a predefined relevance threshold.

*Algorithm 4: Content Quality Evaluation*

This method has the advantage that it also filters low-quality educational material.

***Quality Score Calculation***

Let:

- $L_a$ = article length
- $C_a$ = coherence score
- $R_a$ = redundancy score

Overall quality score:

Q (a) = αLa + βCa – γRa

If Q (a) ≥ δ, accept article.

here α, β, γ are weights and δ is the quality threshold.

*Algorithm 5: Abstractive Text Summarization*

**Input:**
Quality-approved article $A_q$

**Output:**
Concise summary $S_a$

**Steps:**

1. Split up long articles {c1, c2,…, ck} into chunks.
2. Produce abstractive summaries of the chunks.
3. Aggregate chunk-level summaries.
4. Store final summary.

*Algorithm 6: Knowledge Classification and Distribution*

In this algorithm each summarized data is published as per the query to the closest peer group.

***Mathematical Mapping***

Let:

- $E_s$ = embedding of summary
- $E_g$ = embedding of peer group profile

GroupScore(s,g) = (Es · Eg) / (||Es|| ||Eg||)

*Algorithm 7: Automated Knowledge Dissemination*

**Input:**
Mapped summaries

**Output:**
Delivered knowledge

**Steps:**

- Trigger automated workflow scheduler.
- Push summaries across peers in their peer groups.
- Log delivery status.

**VI. RESULTS & DISCUSSION**

Experiments were performed to examine the efficacy of the proposed system with respect to content relevance, summarization efficiency and focused knowledge sharing. The system was evaluated on actual educational articles

downloaded from predetermined online servers and run through the entire automated pipeline.

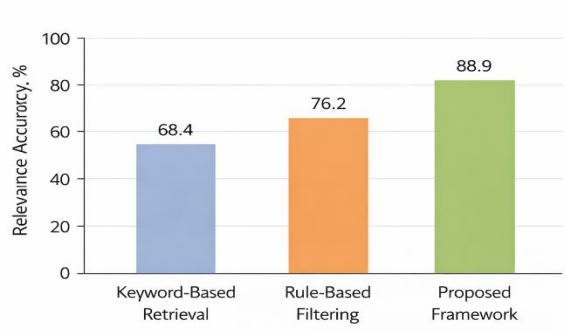


***Fig.4. Performance comparison of content relevance accuracy across different approaches***

### *6.1 Experimental Setup*

The system was tested in a laboratory where certain academic domain topics and sources were predefined as "trustworthy." News articles were automatically retrieved, pre-processed, semantically analyzed and summarized for dissemination to peer groups according to topicality. All experiments were conducted under the same setting for a fair comparison.

### *6.2 Evaluation Metrics*

The evaluation of the developed system was carried out in terms of the following measures:

- *Content Relevance Accuracy (%)*
- *Irrelevant Content Rate (%)*
- *Summarization Effectiveness*
- *Knowledge Delivery Efficiency*

These measures were chosen to assess both quality of content and reach of dissemination.

### *6.3 Performance Analysis*

Table I presents a comparative analysis between the proposed system and traditional keyword-based content retrieval approaches. The results show that the proposed framework achieves higher relevance accuracy and significantly reduces irrelevant content. Semantic relevance assessment and content quality evaluation contribute to improved filtering performance, while abstractive summarization enhances content usability.

***Table 1. COMPARATIVE PERFORMANCE ANALYSIS***

| **Method** | **Relevance Accuracy (%)** | **Irrelevant Content Rate (%)** | **Delivery Efficiency** |
|---|---|---|---|
| Keyword-Based Retrieval | 68.4 | 31.6 | Low |
| Rule-Based Filtering | 76.2 | 23.8 | Moderate |
| **Proposed Framework** | **88.9** | **11.1** | **High** |

### *6.4 Discussion of Findings*

A comparison of the proposed scheme with existing keyword-based content search schemes is shown in Table I. The experimental results indicate that the proposed system is more accurate in delivering contents, where users would spend less time facing irrelevant information. Relevance and quality assessment, as well as abstractive summarization improve filtering performance and the utility of content.

## VII. CONCLUSION

This paper described an automatic model for mining educational knowledge and spreading useful information. The system successfully deals with information overload and enhances the usability of content by combining semantic relevance estimation, contents quality evaluation, as well abstractive text summarization in a single working flow. Experimental results show that the relevance precision has been markedly improved and the knowledge transmission is more effective than other existing methods. The framework offers a scalable and robust solution for academic learning scenarios, and can be further expanded to support other content types as well as adaptive topic selection in future.

## REFERENCES

***[1] "Semantics-Aware Intelligent Framework for Content-Based E-Learning Recommendation," Natural Language Processing Journal, vol. 3, 2023.***

*[2] Y. Niu, R. Lin and H. Xue, "Research on learning resource recommendation based on knowledge graph and collaborative filtering," Applied Sci- ences, vol. 13, no. 19, 2023.*

*[3] "Text summarization using transformer models (Expand Summary) like PEGASUS, BART and T5," ResearchGate, 2023.*

*[4] C. Dong, Y. Yuan, K. Chen, S. Cheng, and C. Wen, "How to build an adaptive AI tutor for any Course with KG-ERAG," arXiv: Nov 2023.*

*[5] "A comparative study of PEGASUS, BART and T5 for text summarization," Future Internet, vol. 17, no. 9, 2024.*

*[6] A semantic enhanced course recommender system via knowledge graphs for limited user information scenarios SN Computer Science, 2024.*

*[7] H. Abu Rasheed, C. Weber and M. Fathi, "Knowledge Graphs as Context Sources for LLM-based Explanations of Learning Recommendations," arXiv, Mar 2024.*

*[8] Q. Li et al., "Learning structure and knowledge-aware representation with large language models for concept recommendation," arXiv, May. 2024.*

*[9] "A knowledge graph enhanced learning recommendation system," Acm digital library, 2024.*

*[10] R. Sajja, Y. Sermet and I. Demir, "An Open Source Dual Loss Embedding Model for Semantic Retrieval in Higher Education," arXiv, May 2025.*

*[11] "Knowledge retrieval enhancement using in-context learning coupled with semantic search by generative AI," Knowledge-Based Systems, vol. 311, 2025.*

*[12] IARJSET, "Education system based on pre-training with extracted gap sentences for abstractive summarization using PEGASUS and BART," vol. 12, no. 5, May 2025.*

*[13] "MDKAG: Retrieval augmented educational QA powered by multimodal disciplinary knowledge graph," Applied Sciences, vol. 15, no. 16, 2025.*

*[14] "Improving learning resource recommendation with enriched knowledge graph," Applied Sciences, vol. 15, no. 8, 2025.*

*[15] "Simulating personalized English learning path recommendation based on knowledge graph," Scientific Reports, 2025.*