

## Article

# From Profiles to Promising Paths: A Semantic Group Recommender for Novel Academic Topic Discovery

Carlos Ayala-Tipan <sup>1,\*</sup> , Lorena Recalde <sup>1</sup>  and Edison Loza-Aguirre <sup>2</sup> 

<sup>1</sup> Department of Informatics and Computer Science, Escuela Politécnica Nacional, Quito 170525, Ecuador; lorena.recalde@epn.edu.ec

<sup>2</sup> Colegio de Administración de Empresas, Universidad San Francisco de Quito, Quito 170901, Ecuador; eloza@usfq.edu.ec

\* Correspondence: carlos.ayala01@epn.edu.ec

## Abstract

Scientific production is expanding so quickly that research teams are struggling to track advances beyond their immediate specialization, especially in interdisciplinary areas where relevant work is scattered across venues and vocabularies. To reduce this overload at the group level, we propose an end-to-end pipeline that transforms structured bibliographic metadata into actionable topic recommendations for research teams. Starting from Scopus records, the method normalizes scholarly text, builds semantic author profiles using Sentence-BERT representations coupled with interpretable keyword descriptors, and forms candidate groups from co-authorship signals and profile similarity. For each group, the approach applies embedding-based topic modeling to generate candidate themes and ranks them using a relevance–novelty trade-off, enabling teams to surface directions that remain aligned with their collective agenda while still encouraging exploration beyond dominant or highly popular topics. Empirical evidence on a Scopus-derived corpus shows that embedding-aware descriptors support cleaner, semantically faithful representations than frequency-based baselines, strengthening downstream topic discovery and producing compact topic lists that are easier for teams to inspect, discuss, and adopt in collaborative planning.

**Keywords:** academic profiling; BERTopic; group recommender systems; KeyBERT; semantic embeddings; SentenceBERT

## 1. Introduction

Scientific production is growing at a pace that makes it increasingly difficult for researchers to remain aware of relevant advances beyond their immediate specialization. This phenomenon is particularly visible in interdisciplinary areas, where the same research problem can be addressed through different methodological traditions and venues. In practice, researchers face an information overload scenario: the number of candidate papers, venues, and emerging topics exceeds the time and cognitive capacity available to systematically explore them [1,2]. Figure 1 illustrates this imbalance, where a large number of potential sources competes for limited attention, motivating the need for computational support to filter, prioritize, and guide exploration.

Recommender systems (RS) provide a natural response to this challenge by ranking items (e.g., papers, venues, or topics) according to user preferences and historical signals [3]. However, academic decision-making frequently occurs in group settings rather than in



Academic Editors: Costas Vassilakis and Dionisis Margaritis

Received: 28 April 2026

Revised: 8 June 2026

Accepted: 15 June 2026

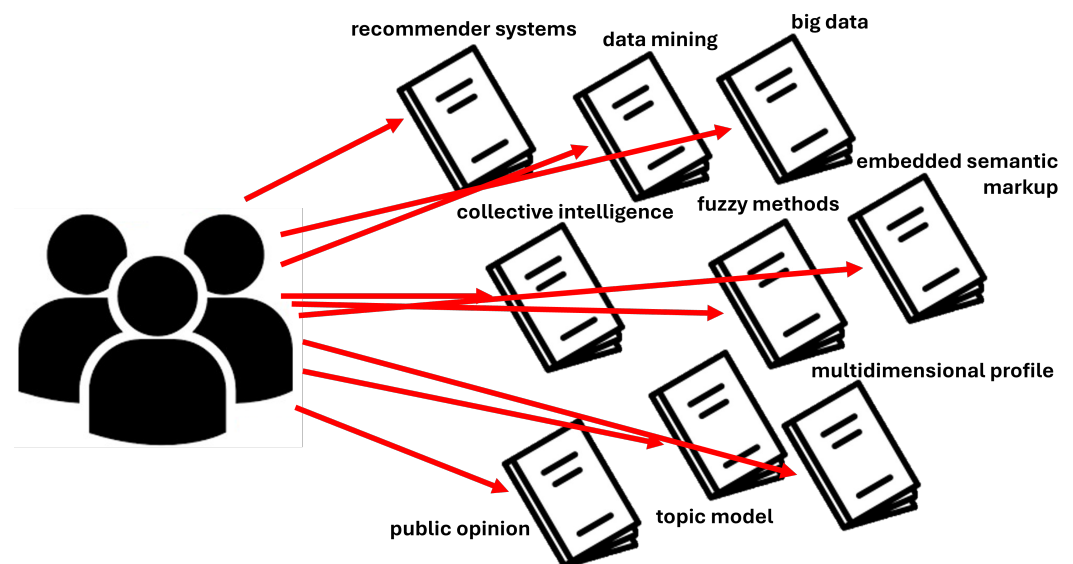
Published: 22 July 2026

**Copyright:** © 2026 by the authors.

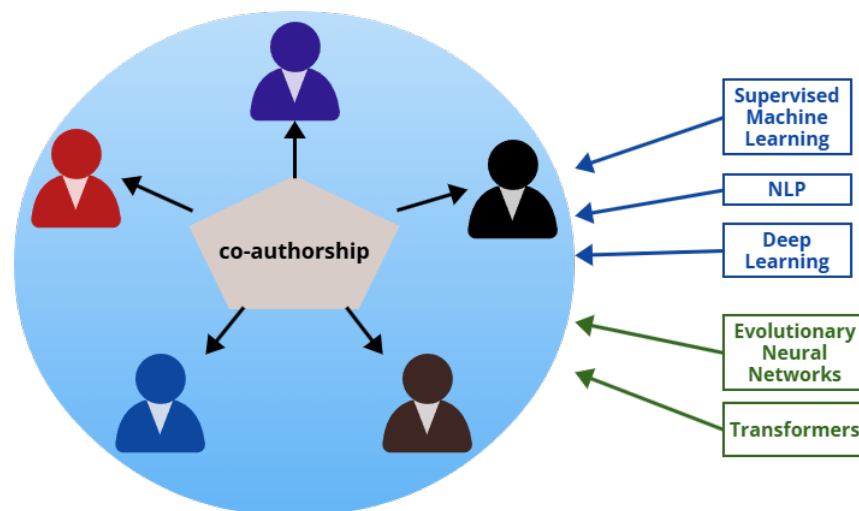
Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

isolation. Research groups, laboratories, and collaborations must collectively decide which topics to explore, which directions to prioritize, or which methodological lines to adopt. In these contexts, the goal is not only to satisfy individual interests but to support a shared research agenda while preserving fairness and representativeness among members. Group recommender systems (GRS) address this setting by aggregating multiple preferences into a collective recommendation, providing benefits such as coordinated exploration and reduced negotiation overhead [4,5]. Figure 2 summarizes these motivations and highlights why a group-oriented perspective is essential when recommendations are intended to support collective research planning.



**Figure 1.** Information overload in academic settings motivates computational support to guide exploration and reduce the cost of identifying relevant directions.

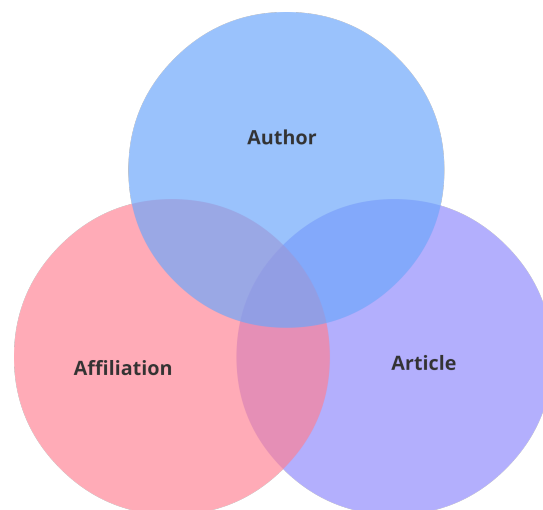


**Figure 2.** Group recommender systems support collective decision-making by aggregating preferences and coordinating exploration, which is particularly relevant for research groups and laboratories.

Despite their potential, deploying GRS in academic environments involves challenges that are often underestimated. First, academic interests are high-dimensional and dynamic: topic preferences evolve with new projects, collaborations, and funding opportunities. Second, bibliographic data are typically sparse and long-tailed: many authors have few publications, and a small subset of authors accounts for a large portion of the content, which complicates the construction of robust user representations and stable group profiles. Third,

recommendations for research planning should not merely reproduce popular topics; they should help groups identify directions that are simultaneously relevant and informative while encouraging intellectual exploration. This requirement brings additional quality dimensions into play, including novelty and diversity, and it also echoes the broader recommender systems finding that optimizing relevance alone can be insufficient for user and stakeholder objectives [6,7].

A key enabler for academic GRS is the availability of structured bibliographic metadata. Modern scholarly databases expose information that connects authors, affiliations, and publications, allowing systems to infer research profiles and collaboration signals from observable traces [8]. Figure 3 presents the central entities and their relationships, which underpin the data pipeline in this study. In particular, we leverage Scopus metadata accessed through an API-driven process and transform it into representations that support (i) profile modeling at the author level, (ii) group formation based on collaboration signals and profile similarity, and (iii) topic recommendation at the group level [9]. In this study, we instantiate the pipeline on a country-scoped scholarly snapshot by filtering Scopus records to authors affiliated with Ecuador, which provides a concrete setting to analyze collaboration structures and topic discovery under realistic metadata constraints.



**Figure 3.** Core bibliographic entities and relations used in this study (Author–Affiliation–Article), enabling profile modeling, collaboration signals, and topic recommendation.

In recent years, some work has addressed academic profile modeling through bibliographic traces and keyword-centered representations, showing that scholarly metadata can support meaningful researcher characterization [10]. In this article, we address the broader problem of group topic recommendation in academic settings. The proposed solution combines semantic representations, keyword extraction, collaboration signals, and topic modeling in a unified end-to-end pipeline that produces interpretable recommendations for research teams. Beyond the algorithmic design, we provide an evaluation that compares multiple baselines across controlled scenarios, allowing the contribution of each modeling choice to be discussed with clarity.

The main contributions of this work are as follows:

- We propose an end-to-end framework for academic group topic recommendation built on bibliographic metadata, connecting data acquisition, preprocessing, profile modeling, group formation, and topic recommendation in a coherent pipeline.
- We operationalize semantic profile modeling as an intermediate representation for academic group topic recommendation, adapting transformer-based document representations and interpretable keyword descriptors to connect individual biblio-

graphic evidence, group formation, topic candidate generation, and relevance–novelty-aware ranking.

- We operationalize quality dimensions relevant to research planning—including relevance, novelty, and diversity—and incorporate them into the recommendation narrative and experimental analysis.
- We conduct a comparative evaluation across multiple scenarios and baselines, reporting quantitative and qualitative evidence to support the effectiveness of the proposed approach.

The remainder of the paper is organized into five sections beyond this introduction. Section 2 reviews related work on scholarly recommender systems, group recommendation, and topic modeling. Section 3 presents the proposed approach, detailing the end-to-end pipeline from data acquisition and preprocessing to profile modeling, group formation, and topic recommendation. Section 4 describes the experimental setup, including the dataset scope, scenario definitions, and the hyperparameters used for reproducibility. Section 5 reports the results and discusses insights and limitations under the relevance–novelty–diversity perspective. Finally, Section 6 concludes the paper and outlines future research directions.

## 2. Related Works

Scholarly recommender systems have evolved from classic information retrieval and recommender paradigms toward richer representations that combine text semantics, citation/co-authorship graphs, and heterogeneous metadata. In academic contexts, the goal is not only to rank papers or venues, but also to support researchers in identifying topics that are relevant to their current agenda while still expanding their intellectual horizon. This objective becomes increasingly difficult as publication volumes grow, signals concentrate on already-visible authors and venues, and exposure effects amplify popularity bias [11–14]. Consequently, modern pipelines tend to combine multiple evidence sources (text, citations, authorship, venues, affiliations) and explicitly address beyond-accuracy objectives such as novelty, diversity, and fairness [3,7,15,16].

This section positions our work at the intersection of (i) scholarly and academic recommender systems, (ii) group recommender systems (GRS) for collective decision support, (iii) profile modeling and interpretable representations from scientific corpora, and (iv) novelty/diversity-aware recommendation under long-tail and popularity effects. While related works often treat these components independently, our pipeline integrates them to support topic-level discovery for research teams.

### 2.1. Scholarly Recommender Systems

Early scholarly recommendation approaches largely mirrored traditional recommender families: content-based methods leverage textual similarity (titles, abstracts, keywords) and metadata matching (venue, domain categories), while collaborative filtering exploits historical interactions such as clicks, downloads, libraries, or co-reading patterns. Hybrid models combine both to mitigate sparsity and cold-start issues [17,18]. Surveys in paper recommendation and digital-library recommenders report that, in real scholarly settings, explicit feedback is limited and often proprietary, pushing many solutions toward publication metadata and graph signals as the most robust and shareable evidence [11,12].

A defining trait of scholarly data is that it naturally forms a heterogeneous network: authors, papers, venues, affiliations, and citations can be modeled as a multi-relational graph, enabling graph-based inference and representation learning. Recent scholarly recommenders thus increasingly rely on learned representations for papers/authors derived from text encoders and graph context [19]. This shift supports semantic generalization across

terminological variation (e.g., the same theme expressed with different keywords across venues) and reduces dependence on lexical overlap. At the same time, these models inherit well-known recommender challenges: cold-start for early-career researchers, incomplete and noisy metadata, and biases induced by visibility-driven signals such as citation counts and venue prestige [12,14,16].

A practical consequence is that scholarly recommenders are increasingly designed as decision-support systems rather than pure ranking engines. Platforms and deployments in digital libraries highlight the need for reproducibility, transparent logging, and careful evaluation protocols that separate offline improvements from real user impact [6,20,21]. This is particularly relevant when recommendations are used to steer research exploration, where the cost of reinforcing narrow trajectories can be high.

## 2.2. Group Recommender Systems for Research Teams

GRS extends recommendation from individuals to groups by defining how group preferences are elicited, aggregated, and transformed into a final ranked output. Classic strategies include aggregating individual utility scores (e.g., average, least-misery, most-pleasure), rank aggregation variants, and more interactive approaches when members must converge on a joint choice [4,5]. In academic environments, groups are natural decision units—labs, co-authorship communities, and project teams—and recommendations may support collective activities such as selecting a shared reading agenda, identifying funding-aligned topics, or prioritizing research directions.

In the scholarly domain, related recommendation tasks have mainly been studied from adjacent perspectives, such as paper recommendation, collaborator or researcher recommendation, and scholarly resource discovery [22,23]. These works confirm that academic recommendation can extend beyond paper ranking and may support collaboration and expertise discovery. However, they usually target individual users or pairwise collaboration opportunities rather than group-level topic recommendation. Likewise, recent work on content-based group recommender systems provides useful foundations for aggregating content representations at the group level [24], but these approaches are not specifically designed for scholarly teams that require topic-level, interpretable, and novelty-aware recommendations. Therefore, the present work should be understood as an early contribution toward academic group topic recommendation, rather than as a claim that no prior scholarly recommendation systems exist.

A recurring distinction in GRS is whether groups are persistent (e.g., stable labs) or ephemeral (ad hoc collaborations). Ephemeral settings often lack group-level historical interactions and thus rely on aggregating member profiles, whereas persistent settings can be treated as “virtual users” with direct group interaction histories [4,25]. Our problem aligns with the profile-aggregation family, because group decisions are formed from members’ scholarly traces rather than explicit group feedback.

A central challenge in GRS is the trade-off between accuracy (satisfying each member) and collective properties such as fairness, coverage of diverse interests, and avoidance of dominance by a single member—an issue that becomes sharper in long-tail scholarly data. Recent work explicitly formulates fairness-aware group recommendation through Pareto-style or compromise-driven mechanisms that aim to prevent systematic under-representation of minority preferences [25]. These ideas align with our motivation to support collective planning while keeping the recommendation list manageable and representative.

This positioning also affects the evaluation strategy. Since existing academic recommender systems usually focus on paper, venue, expert, or collaborator recommendation, and existing GRS baselines are commonly designed for consumer-oriented domains with

explicit user–item ratings, a direct solution-level comparison is not straightforward for our setting. For this reason, the evaluation in this manuscript focuses on comparable pipeline components: keyword/profile modeling is compared against TF–IDF and CF–IDF baselines, while topic generation and recommendation behavior are analyzed through topic descriptors, novelty scores, and a worked group-level example. We acknowledge that a full end-to-end benchmark for academic group topic recommendation remains an open evaluation challenge.

### 2.3. Profile Modeling from Scientific Corpora

Researcher profile modeling aims to capture a compact and reusable representation of a scholar’s expertise, interests, and potential collaboration areas from observable evidence such as publications, abstracts, keywords, citations, venues, and co-authorship patterns. In early academic recommender and expert-finding settings, profiles were commonly represented through bag-of-words, TF–IDF vectors, manually defined descriptors, or multi-faceted textual profiles extracted from scientific publications [26,27]. These approaches are interpretable and computationally simple, but they can be sensitive to vocabulary mismatch and sparse publication histories, which are common in scholarly datasets.

Recent work has moved toward neural and embedding-based representations for scientific documents and researchers. Citation-informed encoders such as SPECTER learn document-level representations by combining paper text with citation context [19], while contrastive approaches such as SciNCL improve scientific document embeddings by using neighborhood information from citation graphs [28]. At the researcher level, recent studies have also shown that aggregating embeddings from publication titles or abstracts can produce useful researcher representations for academic search, classification, and similar-researcher retrieval tasks [29]. These works show a clear trend: scholarly profile modeling is shifting from purely lexical profiles toward semantic representations that better capture topical similarity across different vocabularies.

At the same time, profile modeling in academic environments must remain interpretable. Research teams typically need topic descriptors that can be inspected, discussed, negotiated, and connected to actionable activities such as literature exploration, grant planning, or proposal writing. For this reason, latent vector representations alone are not sufficient for our setting. The profile component used in this manuscript combines semantic document embeddings with keyword/keyphrase descriptors, so that each profile can be used computationally for similarity operations and also exposed to users as a readable representation of research interests.

This positioning clarifies the scope of our contribution. We do not claim to introduce a new embedding architecture or a new standalone researcher profiling algorithm. Rather, the contribution lies in adapting semantic profile modeling to the academic group recommendation domain and in providing empirical evidence of its role as an intermediate representation within the complete recommendation pipeline. In the proposed framework, semantic author profiles serve as a functional bridge between individual bibliographic evidence, co-authorship-based group formation, topic candidate generation, and relevance–novelty-aware ranking. Previous work has shown that bibliographic traces and keyword-centered representations can support academic profile modeling [10]. In this study, this idea is operationalized at the group level, where the value of the profile is assessed by how it supports interpretable and novelty-aware topic recommendations for research teams.

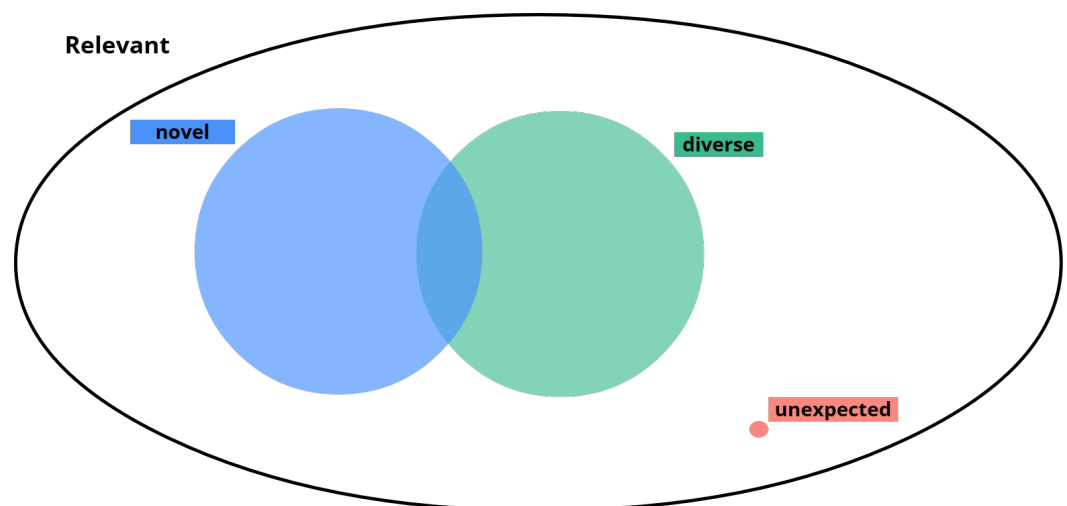
### 2.4. Novelty, Diversity, and Popularity Bias in Scholarly Recommendation

Beyond relevance, modern recommendation research emphasizes objectives such as novelty, diversity, serendipity, and exposure fairness. The motivation is particularly

strong in scholarly scenarios: if ranking is optimized only for similarity to past interests, systems may reinforce narrow research paths and amplify already popular topics [13,14,16]. Long-tail dynamics are not only a statistical feature of item consumption but also an exposure mechanism: users interact with what is recommended, and recommended items become more visible, creating feedback loops that can further marginalize emerging or niche topics [16].

Operationalizing these objectives requires distinguishing multiple notions of “interestingness”. Relevance ensures thematic alignment with the group’s current profile; novelty promotes less-seen but still pertinent themes; diversity encourages coverage across different subareas; and unexpectedness (often associated with serendipity) captures recommendations that are not trivially inferred from the dominant profile yet remain justifiable. These objectives are well-established in beyond-accuracy recommendation research [3,7,15] and provide a principled basis for designing scholarly recommenders that support exploration rather than only exploitation.

Figure 4 frames the optimization problem that we address: the system must remain inside the “relevant” region while shifting from purely familiar topics toward a controlled mixture that also increases novelty and diversity. This conceptual framing bridges related work and our proposed method because it clarifies why a purely relevance-driven baseline is insufficient for group-level topic discovery in research planning.



**Figure 4.** Conceptual relationship between relevance, novelty, diversity, and unexpectedness in the space of recommended topics.

Within the existing literature, recurring limitations emerge that directly motivate the design choices of our proposal. First, many scholarly recommenders emphasize paper-level recommendation, whereas research teams frequently need topic-level suggestions to guide collective planning [11,12]. Second, group recommendation methods are often evaluated in consumer domains; when moved to academic teams, they demand reliable and interpretable profiling, clear topic descriptors, and careful treatment of novelty/diversity objectives to support real collective deliberation [4,5,25]. Third, although profile modeling pipelines and topic modeling components are widely studied, they are commonly presented as separate contributions; comparatively fewer works describe an end-to-end flow that connects metadata ingestion, group formation, topic generation, and final recommendation in a single operational framework. Fourth, popularity bias and long-tail dynamics remain a practical risk in scholarly recommendations, motivating explicit mechanisms to preserve relevance while promoting exploration [13,14,16]. These observations provide the

context in which our pipeline contributes: it integrates profiling, grouping, and topic-level recommendation while explicitly accounting for exploration-oriented quality dimensions.

To make these comparisons explicit, Table 1 maps how the above lines of work relate to the pipeline we implement and where our approach extends them.

**Table 1.** Summary of related work and their limitations for group topic recommendation in academic contexts.

| Line of Work                           | Typical Focus                                                                                                 | Limitation for Our Setting                                                                                                                                                              |
|----------------------------------------|---------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Scholarly recommenders                 | Paper, venue, resource, expert, or collaborator recommendation using text, metadata, and graphs [11,12,22,23] | Often individual- or pairwise-oriented; limited support for team-level topic discovery and explicit novelty/diversity control                                                           |
| Group recommender systems              | Aggregation and consensus to produce a joint ranking; content-based group modeling [4,5,24]                   | Usually evaluated in consumer-oriented domains; academic group topic recommendation requires explainable scholarly profiles and domain-specific objectives (novelty/diversity/fairness) |
| Profile modeling & keyword extraction  | Author/document representations; embedding-driven descriptors [10,19]                                         | Frequently stops at profiles; does not close the loop into group-level topic recommendation                                                                                             |
| Novelty/diversity-aware recommendation | Mitigating popularity bias; promoting exploration [14–16]                                                     | Needs to be integrated with relevance constraints and validated in scholarly/team scenarios                                                                                             |

Taken together, the reviewed literature suggests that existing approaches solve only parts of the problem addressed in this article. Scholarly recommenders are often paper-centered, group recommenders usually depend on preference aggregation without strong semantic profiling, and profile modeling studies frequently stop before reaching topic-level recommendation for collaborative teams. As a result, current solutions offer limited support for research groups that need recommendations that are simultaneously interpretable, semantically grounded, and open to novelty.

This makes an end-to-end strategy particularly appropriate in our setting, since it connects bibliographic data acquisition, researcher profile construction, group formation, and topic recommendation within a single operational framework. By integrating these stages, the proposed approach is better positioned to preserve semantic fidelity, reduce the effect of noisy or sparse metadata, and generate group-level topics that are not only relevant, but also manageable and actionable for collective research planning.

### 3. Proposed Approach

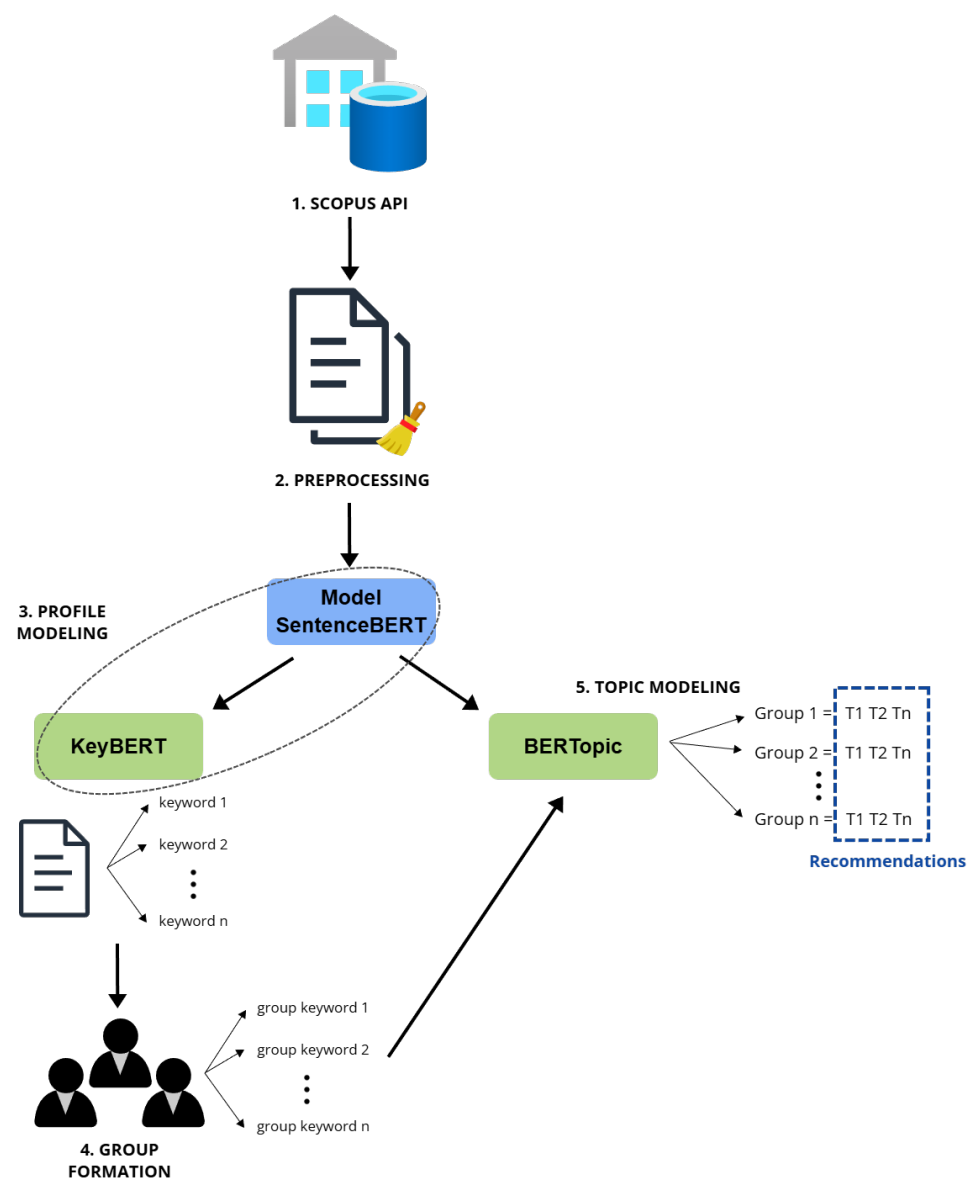
To address the limitations identified in the related literature, the proposed approach is organized as a coherent end-to-end pipeline that transforms raw bibliographic metadata into actionable, group-level topic recommendations. Rather than presenting isolated components, the method is described as a sequence of stages: data acquisition, preprocessing, profile modeling, group formation, topic modeling, and ranking.

Where each stage produces an explicit artifact (datasets, indices, embeddings, topics, or ranked lists) that constrains and strengthens subsequent steps. This structure is intentionally aligned with academic team workflows: researchers and laboratories typically need interpretable representations (profiles and topics) and practical outputs (short ranked lists) that can be discussed and adopted collectively.

### 3.1. Design Goals and Overview

The approach targets a recurring problem in academic environments: forming coherent groups of researchers and recommending novel yet relevant research topics while reducing the burden of information overload. Figure 5 summarizes the main stages: (i) Scopus API ingestion, (ii) preprocessing, (iii) profile modeling, (iv) group formation, and (v) topic modeling and recommendation.

Three design goals guide the pipeline. First, semantic representation quality: represent authors and documents with semantic embeddings to reduce vocabulary mismatch and capture latent research themes [30,31]. Second, group-awareness: shape recommendations at the group level by leveraging co-authorship signals and aggregated topical interests rather than individual-only profiles [32]. Third, scalability under long-tail data: introduce filtering, caching, and index structures so the pipeline remains feasible when author productivity is strongly skewed (cf. Figure 7).



**Figure 5.** End-to-end pipeline of the proposed approach (Scopus ingestion, preprocessing, profile modeling, group formation, topic modeling, and recommendations).

### 3.2. Data Model and Acquisition

The pipeline consumes bibliographic records collected through Elsevier Scopus APIs, which provide structured metadata for authors, documents, and affiliations [33]. Each record is mapped into three tightly connected entities: Author, Article, and Affiliation. This explicit separation avoids conflating an author's identity with institutional mobility and supports later steps such as disambiguation, filtering by institution/country, and group formation. Figure 3 reflects this conceptual model and motivates the minimum set of fields required for downstream processing (e.g., Scopus author ID, title/abstract/keywords, year, venue, affiliation identifiers).

Seed definition and retrieval strategy.

Given an initial seed list of researchers defined by an affiliation-based scope (e.g., institutional rosters, laboratory lists, or known Scopus author identifiers), the system queries Scopus to retrieve publication metadata and builds a unified dataset. Two practical retrieval patterns are supported: (i) author-centric, where the pipeline iterates over known author IDs and retrieves associated documents; and (ii) affiliation-centric, where affiliation constraints (institution, country, or organizational scope) are used to discover authors and then expand to their publications. Both patterns store raw API payloads and normalized tables to preserve traceability.

Integrity constraints and de-duplication.

Academic metadata may contain duplicates (e.g., the same document appearing under multiple retrieval paths) and incomplete records (e.g., missing abstracts). The acquisition stage applies integrity checks such as: duplicated EIDs/DOIs, inconsistent author identifiers, empty titles, and missing abstract fields. Records that pass constraints are stored with standardized identifiers (EID/DOI) and linked to authors and affiliations for later aggregation.

### 3.3. Preprocessing and Text Normalization

Preprocessing transforms heterogeneous bibliographic text into a stable input for embedding, keyword extraction, and topic modeling. Titles and abstracts are concatenated into a document representation, optionally enriched with author keywords and venue descriptors when available. The text is normalized by: (i) lowercasing, (ii) removing markup artifacts, (iii) collapsing repeated whitespace, (iv) standardizing punctuation and hyphenation, and (v) optionally filtering extremely short abstracts that are unlikely to represent topical content.

A domain-preserving strategy is applied: acronyms, alphanumeric tokens, and technical terms are retained whenever they carry semantic meaning in scientific writing (e.g., "IoT", "JSON-LD", "OLAP"). This is important because aggressive token filtering can remove exactly the terms that discriminate research areas.

Auxiliary indices for downstream stages.

Two indices are built at this stage:

- Co-authorship index: a graph-oriented structure where each author is connected to co-authors observed in retrieved publications. Edge weights can encode frequency of collaboration and optionally recency (e.g., exponential decay by publication year) [32].
- Document-to-author map: an efficient mapping from each document to its author set, enabling fast aggregation of documents per author and per group without repeated joins.

### 3.4. Researcher Profile Modeling

The core idea of profile modeling is to represent each researcher with a compact, semantically meaningful signature of their research interests. In this manuscript, profile modeling is not presented as a new embedding model, but as an operational intermediate representation adapted to the academic group topic recommendation task. This representation allows individual bibliographic evidence to be connected with group formation, topic discovery, and ranking [10]. In practice, each author profile is learned from the set of documents authored by that researcher, combining semantic representations to reduce lexical mismatch with keyword-based summaries for interpretability.

Semantic document representation.

Let  $d$  be a document constructed from title and abstract (and optionally keywords). A Sentence-BERT encoder produces a dense vector  $\mathbf{e}_d \in \mathbb{R}^m$  [31]. This representation inherits its contextual language understanding from BERT-style pretraining [30] and supports similarity comparisons across different terminologies that refer to the same scientific concept.

Keyword extraction for interpretability.

Profiles must remain inspectable by researchers. To keep profiles interpretable, the pipeline applies an embedding-based, unsupervised keyphrase extraction strategy (implemented with KeyBERT-style scoring) that ranks candidate phrases by cosine similarity between contextual phrase embeddings and the document embedding [30,31,34]. For an author  $a$ , the system builds a weighted multiset  $\mathcal{K}_a$  from the author's documents, where each candidate keyword inherits a relevance score and can be re-weighted by frequency or recency. This produces a human-readable view of interests and supports comparisons with traditional baselines (e.g., TF-IDF, CF-IDF) as presented in the Results section.

Author profile aggregation (semantic core).

Given the set of documents  $D_a$  of author  $a$ , the semantic author embedding  $\mathbf{p}_a$  is computed as a weighted centroid:

$$\mathbf{p}_a = \frac{\sum_{d \in D_a} w_d \mathbf{e}_d}{\sum_{d \in D_a} w_d}, \quad (1)$$

where  $w_d$  can depend on publication recency (to emphasize current interests), venue type, or be uniform when such metadata is not used. This aggregation is stable under sparse publication histories and yields a single vector suitable for similarity-based grouping and topic alignment.

Profile artifacts exposed to users.

For each author, the pipeline maintains: (i) the embedding  $\mathbf{p}_a$  (used computationally), (ii) the keyword set  $\mathcal{K}_a$  (used interpretably), and (iii) the supporting document list (used for traceability). This trio is critical in academic settings: teams can validate whether the profile is faithful before acting on recommendations.

### 3.5. Group Formation via Co-Authorship Signals

The system forms candidate research groups by exploiting co-authorship as a social signal that naturally encodes collaboration and shared interests [32]. Co-authorship is particularly suitable in institutional contexts because it reflects real working relationships, reduces the need for explicit preference elicitation, and supports explainability ("this group is formed from observed collaborations").

Two complementary strategies are supported:

- Structure-driven groups: groups are derived from co-authorship components or dense collaboration neighborhoods, producing naturally cohesive teams.

- Seed-driven groups: starting from a seed author, the pipeline selects the top- $k$  closest co-authors according to cosine similarity over author embeddings:

$$\text{sim}(\mathbf{p}_a, \mathbf{p}_b) = \frac{\mathbf{p}_a^\top \mathbf{p}_b}{\|\mathbf{p}_a\|_2 \|\mathbf{p}_b\|_2}. \quad (2)$$

Cosine similarity was selected because author profiles are represented as dense semantic embeddings obtained from Sentence-BERT-style representations, where vector orientation is commonly used to compare semantic proximity [31]. This is appropriate for the present setting because researchers may have different numbers of publications and different amounts of textual evidence; therefore, similarity should emphasize the direction of the profile vector rather than its magnitude. Alternative measures, such as Euclidean distance, Manhattan distance, or correlation-based similarity, could also be used, but they may induce different neighborhood structures and consequently produce different seed-driven groups. In this study, cosine similarity is used as a standard and interpretable choice for embedding-based semantic comparison, while a systematic sensitivity analysis over alternative similarity measures is left for future work.

The second strategy is useful when a laboratory wants recommendations around a principal investigator or a project lead while still grounding membership in collaboration evidence.

### 3.6. Group Profile Construction

Once a group  $g$  is formed with members  $a \in \mathcal{A}_g$ , the pipeline builds a group representation that captures shared interests while controlling dominance effects caused by productivity skew. A direct centroid provides the simplest common-interest representation:

$$\mathbf{p}_g = \frac{1}{|\mathcal{A}_g|} \sum_{a \in \mathcal{A}_g} \mathbf{p}_a. \quad (3)$$

To prevent domination by a single prolific author, the aggregation can be re-weighted using a bounded function of publication count (e.g.,  $w_a = \log(1 + |D_a|)$ ), clipping, or per-author normalization. These controls are consistent with long-tail behavior and aim to preserve the contribution of less prolific members.

In addition to the embedding  $\mathbf{p}_g$ , the group inherits an interpretable descriptor by pooling member keyword sets:

$$\mathcal{K}_g = \bigcup_{a \in \mathcal{A}_g} \mathcal{K}_a, \quad (4)$$

with optional re-weighting to promote coverage across members (e.g., limiting the number of keywords contributed by each author).

### 3.7. Topic Candidate Generation with BERTopic

Topic candidates are generated from group-associated documents using an embedding-based topic modeling strategy (implemented with BERTopic) that clusters contextual document embeddings and derives compact topic descriptors [35–37]. For each group  $g$ , the input corpus is the union of documents authored by its members,  $D_g = \bigcup_{a \in \mathcal{A}_g} D_a$ . Operationally, the procedure: (i) embeds documents with a Sentence-BERT encoder [31], (ii) reduces dimensionality with UMAP to stabilize neighborhood structure for clustering [36], and (iii) clusters with HDBSCAN to detect variable-density topic regions [37]. Topic descriptors are then obtained through a class-based TF-IDF-style weighting of terms within each cluster.

This step converts a potentially large pool of publications into a manageable candidate list of themes that can be ranked. It also supports interpretability: each recommended

topic can be presented with representative keywords and exemplar papers, enabling group members to inspect why a topic emerged.

The use of BERTopic in this study is supported by the availability of a sufficiently large scholarly corpus and by the use of title, abstract, and keyword information as document-level evidence. Since BERTopic relies on pre-trained transformer embeddings, the topic generation stage benefits from semantic knowledge learned from large external corpora, which can reduce dependence on exact word co-occurrence and vocabulary overlap [35]. Nevertheless, topic stability may decrease when the available corpus is very small or when documents are extremely short, because there may be insufficient textual evidence to characterize authors, groups, and candidate topics. This limitation is especially relevant in short-text topic modeling, where sparse word co-occurrence is a well-known challenge [38]. Therefore, the proposed methodology assumes that each group has enough textual evidence, preferably titles and abstracts, to support meaningful embedding and clustering.

Practical hyperparameters and reproducibility.

Because topic granularity is sensitive to configuration, the pipeline records parameters per run (e.g., UMAP neighborhood size, HDBSCAN minimum cluster size, KeyBERT n-gram range). Table 2 summarizes common parameters reported to ensure reproducibility across institutions and corpora.

**Table 2.** Key hyperparameters recorded for reproducibility.

| Component                      | Parameters (Value Used)                                                                                                                                    | Role                                                                             |
|--------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------|
| Sentence-BERT                  | encoder checkpoint: paraphrase multilingual MiniLM-L12-v2                                                                                                  | Defines the semantic space for document and author embeddings                    |
| KeyBERT                        | keyphrase n-grams: (1,3); top- $k$ : $k = \text{int}(\text{num\_keywords})$ (per-article); use_maxsum=True; use_mmr=True; nr_candidates=20; diversity=0.65 | Controls keyword specificity, redundancy reduction, and profile interpretability |
| Keyword evaluation (BERTScore) | model_type=bert base multilingual-cased; rescale_with_baseline=True                                                                                        | Semantic comparison between generated and author-provided keywords               |
| Group formation (FP-Growth)    | min_support = $3/\text{len}(\text{df})$ ; group size filter: $ G  \geq 3$                                                                                  | Finds frequent co-authorship patterns and filters valid groups                   |
| UMAP (BERTopic internal)       | Default                                                                                                                                                    | Dimensionality reduction prior to density-based clustering                       |
| HDBSCAN (BERTopic internal)    | Default                                                                                                                                                    | Variable-density clustering; controls topic discovery and noise                  |
| BERTopic                       | language='multilingual'; embedding_model= model_sentence; min_topic_size=10; calculate_probabilities=True; verbose=True                                    | Topic candidate generation and topic descriptor extraction                       |

### 3.8. Ranking: Relevance–Novelty Trade-Off

The ranking stage prioritizes topics that are relevant to the group while injecting novelty to expand research directions. This is modeled as a multi-objective decision process where each topic  $t$  receives a relevance score and a novelty score, followed by a combined ranking.

Relevance: Each topic  $t$  is represented by an embedding  $\mathbf{e}_t$  (e.g., centroid of embeddings of documents assigned to the topic). Relevance is computed as similarity to the group profile:

$$\text{Rel}(g, t) = \text{sim}(\mathbf{P}_g, \mathbf{e}_t). \quad (5)$$

**Novelty:** Novelty is estimated as distance from the group’s current footprint. One operationalization emphasizes “novel to any individual member”:

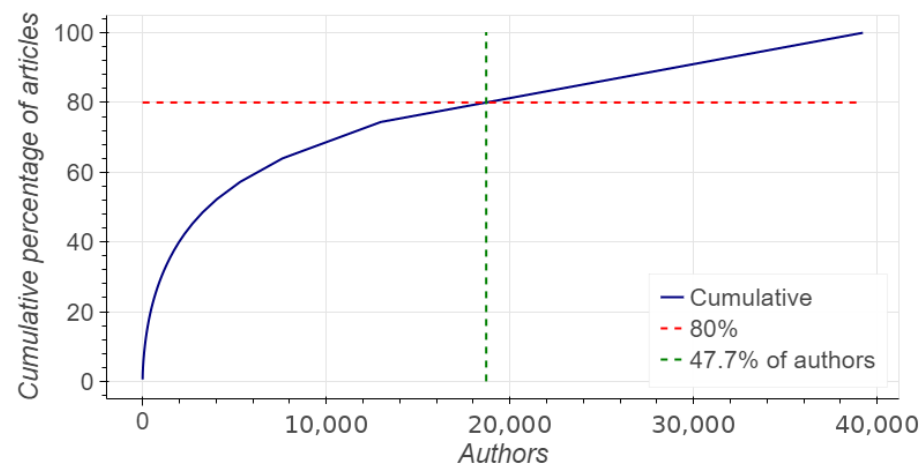
$$\text{Nov}(g, t) = 1 - \max_{a \in \mathcal{A}_g} \text{sim}(\mathbf{p}_a, \mathbf{e}_t), \quad (6)$$

which favors topics that are not overly aligned with a single established profile. A group-centric alternative replaces the max operator with  $\text{sim}(\mathbf{P}_g, \mathbf{e}_t)$  to define novelty relative to the group mainstream.

**Final score and Pareto view:** A simple scalarization ranks topics by a weighted sum:

$$\text{Score}(g, t) = \alpha \text{Rel}(g, t) + (1 - \alpha) \text{Nov}(g, t), \quad (7)$$

where  $\alpha \in [0, 1]$  controls the emphasis on relevance. In parallel, the pipeline can compute the non-dominated set (Pareto front) to expose both conservative and exploratory candidates to users, allowing research teams to choose based on their tolerance for novelty. Figure 6 supports this viewpoint by motivating why cumulative and dominance-aware perspectives are convenient under long-tail behavior.



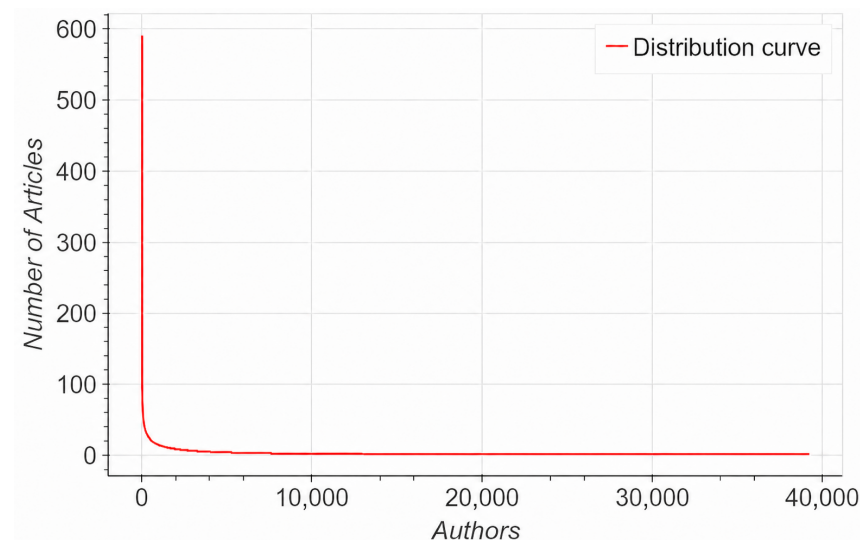
**Figure 6.** Pareto-style cumulative behavior used to motivate filtering/selection under long-tail author distributions and to support multi-objective selection.

### 3.9. Scalability Strategy for Long-Tail Author Distributions

Academic datasets typically present a strong long-tail: many authors publish rarely, while a small fraction concentrates most publications. This impacts both runtime (embedding and topic modeling over large corpora) and group construction (unbalanced evidence across authors). The pipeline incorporates two complementary controls.

**Coverage-based filtering:** When building caches or computing global statistics, the system prioritizes authors and documents that contribute to a target coverage threshold (e.g., the smallest subset of authors that explains 80% of documents), while retaining the option to process the remaining tail incrementally. This keeps early iterations practical without permanently excluding sparse authors.

**Incremental caching and vector indexing:** Document embeddings and extracted keywords are cached so new data updates do not trigger full recomputation. For similarity-heavy operations (e.g., nearest co-authors, topic-to-group similarity), the pipeline can optionally rely on vector indexing to accelerate retrieval [39]. Figure 7 motivates these controls by showing sharp concentration near the head of the distribution.



**Figure 7.** Long-tail behavior in the dataset: a small fraction of authors concentrates a large number of publications.

Table 3 describes the approach as modules with explicit inputs and outputs to facilitate replication and ablation studies.

**Table 3.** Modules of the proposed approach with inputs and outputs.

| Module                                  | Input                                                      | Output                                                                       |
|-----------------------------------------|------------------------------------------------------------|------------------------------------------------------------------------------|
| Data acquisition                        | Seed author IDs; Scopus queries; affiliation constraints   | Raw bibliographic records; author–article–affiliation links                  |
| Preprocessing                           | Titles, abstracts, keywords (raw)                          | Normalized text; document-to-author map; co-authorship index                 |
| Profile modeling (preliminary artifact) | Author document sets; SBERT encoder; KeyBERT               | Author embeddings $\mathbf{p}_a$ ; keyword sets $\mathcal{K}_a$              |
| Group formation                         | Co-authorship graph; similarity over $\mathbf{p}_a$        | Groups $\mathcal{A}_g$ and group corpora $D_g$                               |
| Topic modeling                          | Group corpora $D_g$ ; BERTopic parameters                  | Topics $\mathcal{T}_g$ ; topic embeddings $\mathbf{e}_t$ ; topic descriptors |
| Ranking                                 | $\mathbf{P}_g, \mathbf{e}_t$ , relevance/novelty functions | Ranked topic list per group; optional Pareto set                             |

Algorithm 1 provides an implementation-oriented view of the full pipeline.

**Algorithm 1:** Group topic recommendation pipeline

1. Retrieve Scopus records for the seed author set; build Author–Article–Affiliation entities.
2. Normalize text fields; create document-to-author mapping and co-authorship index.
3. Compute Sentence–BERT embeddings for documents; cache embeddings.
4. For each author, aggregate document embeddings into  $\mathbf{p}_a$ ; extract  $\mathcal{K}_a$  via KeyBERT.
5. Form groups using co-authorship evidence and/or similarity over  $\mathbf{p}_a$ .
6. For each group  $g$ , run BERTopic over  $D_g$  to obtain topics  $\mathcal{T}_g$  and topic embeddings  $\mathbf{e}_t$ .
7. For each topic  $t \in \mathcal{T}_g$ , compute  $\text{Rel}(g, t)$  and  $\text{Nov}(g, t)$ ; rank by  $\text{Score}(g, t)$  or keep Pareto-optimal topics.

## 4. Experimental Setup

This section summarizes the experimental setup used to instantiate and evaluate the proposed pipeline. The goal is to make the evaluation traceable and reproducible, while keeping the narrative consistent with the end-to-end flow described above. All experiments are conducted on a Scopus-derived corpus filtered through a country-level affiliation scope, instantiated in this study with Ecuador-affiliated authors. Ecuador is used here as a realistic and reproducible case-study setting rather than as an exceptional case: it provides a concrete national-scale scholarly subset with heterogeneous bibliographic metadata and observable collaboration patterns, while the proposed pipeline remains applicable to other countries or institutional scopes defined through the same retrieval strategy.

### 4.1. Dataset Scope and Main Entities

For the experiments reported in this article, the pipeline was instantiated on a country-scoped subset obtained by filtering Scopus records to authors affiliated with Ecuador. This case-study configuration provides a realistic setting for analyzing collaboration structures, profile modeling behavior, and group-level topic recommendation under constrained and heterogeneous bibliographic metadata. The dataset is built from bibliographic records retrieved via the Scopus APIs. After integrity filtering and preprocessing, the corpus includes articles with valid identifiers and sufficient textual fields (title and abstract) to support embedding and keyword extraction.

The main entities are: (i) authors (identified by Scopus author IDs and affiliation metadata), (ii) articles (identified by Scopus EIDs/DOIs when available), and (iii) affiliations. The dataset exhibits a long-tail behavior in author productivity (later quantified in Section 5), which motivates the use of semantic profile representations and careful aggregation strategies.

Dataset statistics. The following corpus-level counts are reported in the Results section and are included here for clarity:

- Articles retrieved from Scopus: 39,952.
- Valid articles after preprocessing: 37,768.
- Articles with missing author keywords (NaN): 5670.
- Valid co-authorship groups after filtering: 17,677.

These dataset characteristics define the empirical scope of the current evaluation. The proposed pipeline is intended for scholarly corpora where documents provide enough textual evidence to construct meaningful semantic profiles, typically through titles, abstracts, and, when available, author keywords. If the corpus is much smaller or composed mainly of very short records, the group profile and topic modeling stages may become less stable because author expertise and group interests cannot be characterized with the same level of evidence. In such cases, additional enrichment strategies may be required, such as aggregating documents at the author or group level, incorporating external metadata, reducing topic granularity, or applying short-text topic modeling alternatives.

### 4.2. Keyword Ground Truth and Scenario Definition

Author-provided keywords (when present) are used as a practical reference to evaluate whether extracted descriptors capture the main concepts of each publication. Because the number of keywords varies across articles, the evaluation is performed through three scenarios parameterized by  $N$ , where  $N$  denotes the number of keywords originally reported for a given article. The scenarios follow the evaluation protocol described in the study [10] implementation:

- Scenario 1 (Exact Match): compare the top- $N$  extracted keywords against the  $N$  author keywords.

- Scenario 2 (Extended Match): compare the top- $\frac{3N}{2}$  extracted keywords against the  $N$  author keywords.
- Scenario 3 (Double Match): This design evaluates the behavior of the descriptor extraction strategy under different levels of keyword density and controls the trade-off between coverage (recall) and compactness (precision) of the extracted descriptors.

#### 4.3. Hyperparameters and Implementation Choices

Table 2 reports the hyperparameters used in the experiments. In summary, Sentence-BERT is used to encode titles and abstracts into a shared semantic space; a KeyBERT-style extraction strategy is applied with n-grams (1,3) and a per-article top- $k$  equal to the observed keyword count when available; FP-Growth is used to identify co-authorship groups under a minimum co-publication threshold; and BERTopic is applied to group-associated corpora to produce topic candidates and descriptors.

**Illustrative outputs.** To make the pipeline outputs concrete and easy to audit, Table 4 provides qualitative examples of the profiling stage by contrasting (i) author provided keywords and (ii) keywords generated from the profiling module. These examples show that the extracted descriptors typically preserve the main topical intent of the original terms (e.g., core domain concepts) while adding semantically related phrases that can increase coverage when the original keyword field is sparse or missing. In addition, Table 5 reports examples of group-level topic descriptors together with the novelty index used in our ranking and interpretation. The three groups illustrate distinct novelty profiles: high novelty groups are more open to exploratory topics, moderate novelty groups tend to accept a mix of established and emerging themes, and low novelty groups benefit from more conservative, incrementally novel suggestions to preserve acceptance and perceived relevance.

**Table 4.** Example of articles processed for keyword generation.

| Article ID  | Original Keywords                                                                                                                                                                                                       | Generated Keywords                                                                                                                                                                                                                                        |
|-------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 85089188368 | duran, ecuador, mitigation strategies, tropical wet climate, urban heat island, urban weather generator                                                                                                                 | urban heat island, ecuador estimated spatial, stress probability indoors, general increase air, selected samples classified, strategies policy implications                                                                                               |
| 85089801087 | clustering, cybernetic attacks, health data                                                                                                                                                                             | clustering analysis dataset, cybernetic attacks, health institutions usa                                                                                                                                                                                  |
| 85087712833 | environmental analysis, environmental assessment, environmental health, environmental impact assessment, environmental pollution, environmental science, freshwater, plastics, pollution, quito, rivers, tropical andes | pollution rivers plastics, rivers plastics quito, environmental pollution environmental, plastics quito, andes environmental analysis, health environmental impact, quito freshwater, rivers, freshwater tropical andes, impact assessment, quito, health |

**Table 5.** Generated topics and novelty evaluation for groups.

| Group ID  | Generated Topics                                                                                                                                                                                                                                | Novelty Index (%) |
|-----------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------|
| group2924 | death deaths, patients mortality pulmonary, pandemic sars, patients mortality, cpr, mortality pulmonary, lung respiratory niv, mortality respiratory vaccination, pandemic sars cov, covid patients, pneumoniae mortality, spread health deaths | 65.13             |
| group128  | rocks magmatic, eruption volcano magma, eruptions eruptive, volcano events, magma eruptions, volcanic classification, seismic volcano events, subduction rocks, eruptive ash volcanic, volcano magma eruptions                                  | 30.39             |
| group7170 | diversity phylogenetic, taxonomic clade, biodiversity species, morphological taxonomic, species trees, phylogenetic species, genus phylogenetic, forests tree forest, plot species trees, taxonomic genera                                      | 10.75             |

#### 4.4. Evaluation Metrics

The profiling component is evaluated using congruence metrics that quantify how well extracted keywords match the author-provided descriptors: precision, recall, and co-occurrence ratio for TF-IDF/CF-IDF baselines, and cosine similarity in embedding space for the SBERT + KeyBERT variant. In addition, for semantic keyword-level validation, BERTScore is used to quantify similarity between generated and original keyword sets. These metrics capture complementary aspects: exact lexical overlap (useful for controlled comparison) and semantic relatedness (critical when vocabulary mismatch occurs across disciplines and venues).

### 5. Results and Discussion

This section reports the evidence supporting the proposed end-to-end pipeline, following the same logic introduced in Section 3: (i) characterize the Scopus-derived corpus to justify design decisions for scalability and long-tail-aware processing, (ii) validate the profile/keyword modeling block that acts as the “information bottleneck” between raw metadata and group-level topic discovery, and (iii) interpret outcomes under the relevance–novelty–diversity lens established in Figure 4.

#### 5.1. Results

The results are reported with the practical target in mind: producing useful and manageable topic suggestions for research teams, where usefulness requires semantic fidelity and interpretability, and manageability requires reducing the cognitive burden of exploring a large candidate space (cf. Figure 1). All experiments are reported on the Ecuador-affiliated subset obtained through the Scopus affiliation filter described in Section 3.2.

##### 5.1.1. Corpus Characterization and Author Productivity Skew

A first observation is the strong productivity imbalance across authors, a well-known phenomenon in scholarly corpora and collaboration networks, where a small fraction of actors concentrates a large fraction of outputs [32]. In our dataset, this effect appears as a marked long-tail in the number of publications per author. Figure 7 shows that most authors contribute only a few documents, while a small subset accounts for a disproportionately large number of publications. This distributional property directly motivates two key design choices described in Section 3.9: (i) the pipeline must remain stable under sparse author histories (cold or near-cold profiles), and (ii) group aggregation must avoid being dominated by a handful of prolific authors.

To quantify the imbalance, Figure 6 presents a Pareto-style cumulative analysis. The curve indicates that reaching 80% of the accumulated publications requires approximately 47.7% of the author population. This empirical signal supports the rationale for introducing coverage-based filtering and caching: if embeddings and topic candidates are computed naively for every author/document, the computational budget becomes quickly concentrated on the head of the distribution, while the tail remains underrepresented. In contrast, the proposed pipeline explicitly accounts for this skew and preserves the option of incorporating the long tail progressively, which is consistent with real institutional scenarios where profiles are updated incrementally and new researchers appear continuously.

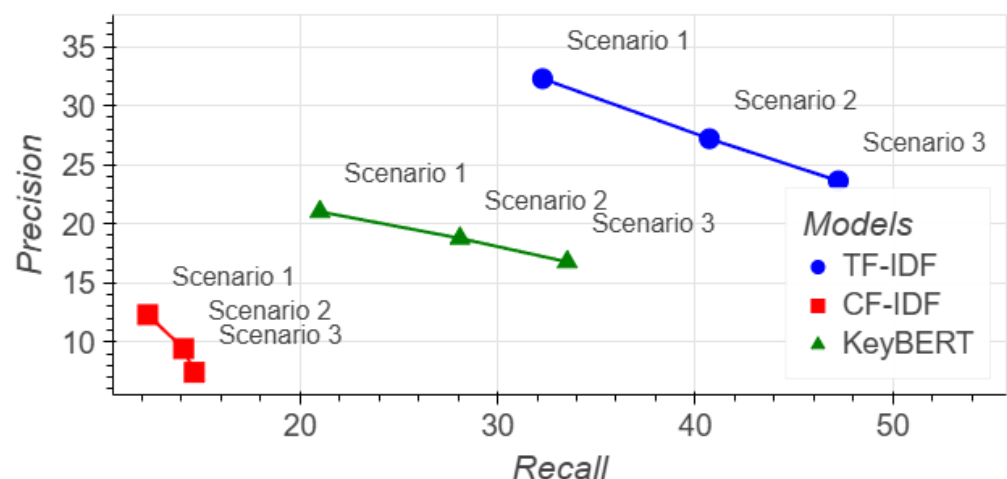
##### 5.1.2. Keyword and Profile Modeling Performance

The pipeline relies on the profile modeling block (Section 3.4) to transform sparse, heterogeneous publication histories into a compact representation used later by group formation and topic recommendation. Concretely, the approach combines semantic document

embeddings (Sentence-BERT) [31] with embedding-aware keyword extraction (KeyBERT) to produce profiles that are simultaneously computable, comparable, and interpretable. This block is evaluated here as a functional component inside the full group topic recommendation pipeline, where the quality of the intermediate profile descriptors affects downstream group formation, topic discovery, and ranking.

To make the scope of this comparison explicit, TF-IDF and CF-IDF are used here as profile/descriptor modeling baselines rather than as full end-to-end group recommendation baselines. This choice allows the evaluation to isolate the contribution of the semantic keyword-based profiling component before its outputs are propagated to group formation and topic recommendation.

Following the experimental protocol, the keyword modeling stage is evaluated across three scenarios (Scenario 1–Scenario 3) against two widely used baselines for keyword/topic descriptor extraction: TF-IDF and CF-IDF. Figure 8 summarizes the precision–recall trade-off across models and scenarios. Two consistent patterns emerge. First, TF-IDF tends to maximize recall, at the expense of precision, retrieving broader candidate sets that often include generic or weakly informative terms. Second, KeyBERT consistently improves the balance: it achieves systematically higher precision than CF-IDF while remaining competitive in recall, yielding a descriptor set that is smaller and easier to validate during group deliberation. CF-IDF shows the weakest behavior in this setting, particularly in recall, which limits its suitability when the objective is to build sufficiently rich candidate pools for later topic discovery.



**Figure 8.** Precision–recall trade-off for keyword descriptor extraction across three scenarios comparing TF-IDF, CF-IDF, and KeyBERT.

Table 6 details the values plotted in Figure 8. TF-IDF reaches the highest recall range (32.30–47.30), while KeyBERT provides a stronger precision range than CF-IDF and maintains moderate recall (20.99–33.54), a desirable compromise when the system must reduce information overload without collapsing the candidate space to only a few overly generic descriptors. Importantly, this effect is not only a matter of keyword quality in isolation: the downstream steps (group profiling and topic modeling) depend on these descriptors to represent topics in a way that remains interpretable for teams, which makes a balanced precision–recall profile operationally more useful than maximizing recall alone.

From a pipeline perspective, the observed behavior aligns with the methodological intent. TF-IDF expands coverage, which can be useful to explore a wide candidate space, but it also introduces lexical noise that complicates group negotiation. KeyBERT, supported by semantic embeddings, filters toward more representative phrases and yields a topic/keyword set that is more suitable as a “decision object” in group recommendation

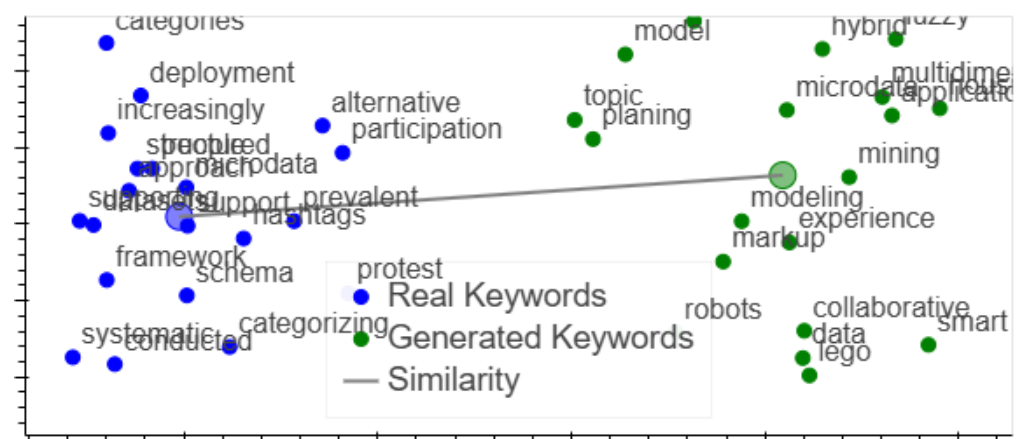
workflows, where recommendations should be inspectable and justifiable rather than only statistically frequent.

**Table 6.** Precision–recall results for keyword modeling baselines and the proposed KeyBERT-based variant across three scenarios.

| Model   | Scenario 1           | Scenario 2           | Scenario 3           |
|---------|----------------------|----------------------|----------------------|
| TF-IDF  | P = 32.34, R = 32.30 | P = 27.24, R = 40.74 | P = 23.81, R = 47.30 |
| CF-IDF  | P = 12.40, R = 12.25 | P = 9.52, R = 14.10  | P = 7.47, R = 14.59  |
| KeyBERT | P = 21.12, R = 20.99 | P = 18.78, R = 28.09 | P = 16.87, R = 33.54 |

### 5.1.3. Semantic Consistency Between Real and Generated Keywords

Precision and recall quantify overlap with reference descriptors, but the pipeline also requires semantic fidelity: extracted descriptors should remain close to the underlying concepts in the corpus, reducing vocabulary mismatch across disciplines and venues. Figure 9 offers a qualitative view of this property by projecting real and generated keywords into an embedding space and highlighting similarity relations. The visualization indicates that generated keywords populate a coherent region and remain connected to the region of real keywords through high-similarity relations, supporting the claim that the keyword modeling step is not producing arbitrary terms but semantically grounded descriptors, consistent with the embedding-driven design [31].



**Figure 9.** Embedding-space visualization of semantic consistency between real (ground-truth) keywords and generated keywords, highlighting proximity relations that support semantic fidelity of the extracted descriptors.

This qualitative evidence complements Table 6: the model retrieves a useful quantity of descriptors (recall), keeps a cleaner subset (precision), and preserves semantic proximity, which is critical for the next stage where group-level topics must be interpreted as meaningful candidates rather than noisy keyword clusters.

### 5.1.4. Component-Level Ablation Analysis

To provide additional insight into the contribution of the main pipeline components, we conducted a component-level ablation analysis focused on the profile and descriptor modeling stage. A full end-to-end ablation is not straightforward in this setting because some modules depend on non-equivalent inputs and outputs: for example, group formation requires author-level representations, while topic recommendation requires group corpora and topic descriptors. Therefore, the ablation analysis focuses on the component where directly comparable alternatives are available: the transformation of bibliographic text into interpretable profile descriptors.

Table 7 summarizes the ablation variants. The lexical variants replace the semantic descriptor module with TF-IDF and CF-IDF, while the full profile descriptor component uses the SBERT + KeyBERT strategy described in Section 3.4. This comparison isolates the effect of using semantic representations and embedding-aware keyword extraction before the descriptors are propagated to group modeling and topic recommendation.

Table 7. Component-level ablation analysis of the profile descriptor modeling stage.

| Variant                             | Ablated or Replaced Component                                                | Observed Effect                                                                                                                                 |
|-------------------------------------|------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------|
| TF-IDF profile descriptors          | Semantic keyphrase extraction is replaced by frequency-based term weighting  | Provides broad lexical coverage and high recall, but introduces more generic descriptors and lower precision                                    |
| CF-IDF profile descriptors          | Semantic keyphrase extraction is replaced by class-based frequency weighting | Produces compact descriptors, but shows the weakest precision–recall behavior across the three scenarios                                        |
| SBERT + KeyBERT profile descriptors | Full semantic descriptor component is retained                               | Produces a better balance between precision and recall than CF-IDF and yields descriptors that are semantically closer to the original keywords |

The results reported in Table 6 support this ablation analysis. TF-IDF reaches the highest recall values, which indicates that frequency-based descriptors can retrieve broader candidate sets. However, this behavior comes at the cost of lower precision and a higher probability of including weakly informative terms. CF-IDF performs worse across the three scenarios, suggesting that class-based frequency weighting alone is insufficient for the heterogeneous scholarly records considered in this study. In contrast, the SBERT+KeyBERT variant provides a more balanced behavior, with higher precision than CF-IDF and moderate recall, making it more suitable for downstream group-level recommendation where descriptors must remain interpretable and semantically meaningful.

This component-level ablation indicates that the semantic descriptor stage is not merely an implementation detail. It acts as an information bottleneck that affects the quality of group profiles and the interpretability of candidate topics. Although a full end-to-end ablation involving all modules is left for future work, this analysis provides evidence that replacing the semantic profiling component with purely lexical alternatives changes the quality and usability of the intermediate descriptors used by the recommendation pipeline.

5.1.5. Interpretation Under Relevance, Novelty, and Diversity Criteria

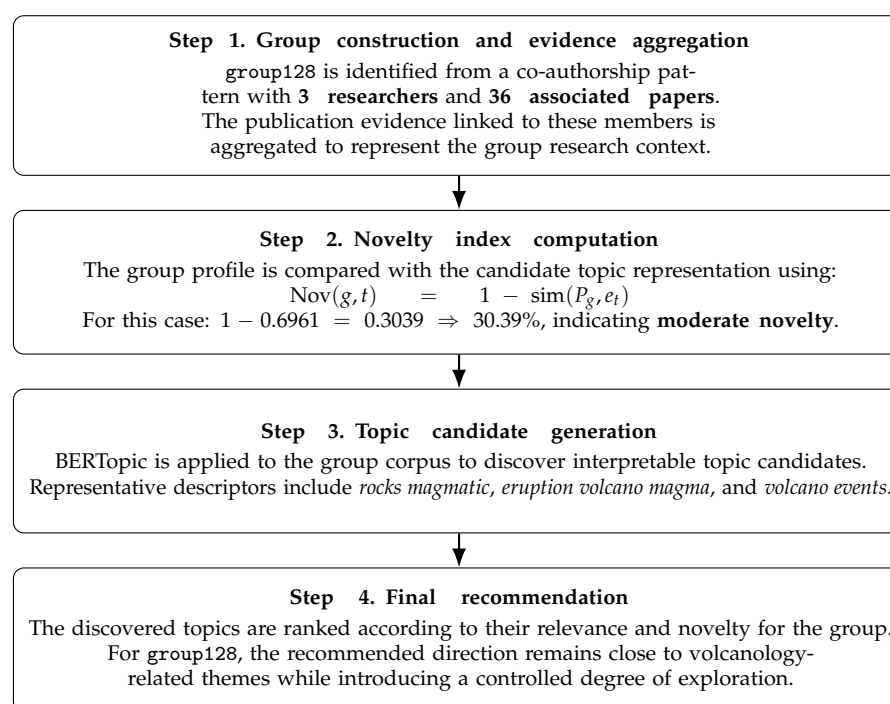
The component-level ablation shows that descriptor quality affects the intermediate representations used by the pipeline. However, the ultimate objective of the proposed system is not to optimize lexical overlap metrics in isolation, but to support group-level topic discovery for research planning. Figure 4 frames this objective through three criteria that matter in collective decision-making: relevance, novelty, and diversity. The results above support relevance directly through improved precision and semantic consistency of descriptors, and they support novelty/diversity indirectly by reducing the tendency of purely frequency-driven representations to collapse onto dominant themes in long-tail corpora [32].

Operationally, when descriptors are semantically faithful and less noisy, the downstream topic modeling stage (BERTopic) can form clusters that are easier to interpret and less likely to be dominated by generic tokens [40]. This increases the feasibility of selecting topics that remain within the relevant region while still moving toward novel and diverse directions, as conceptualized in Figure 4. In practice, this is the key pipeline-level takeaway: improving the quality and semantic structure of profiles is not an auxiliary task, but a

necessary condition for producing group topic recommendations that reduce overload and enable exploration without sacrificing coherence.

#### 5.1.6. Worked Example of the Recommendation Pipeline: Group 128

A compact end-to-end illustration is useful because Tables 4 and 5 show intermediate outputs separately, whereas the following example connects them into a single recommendation trace. To make the end-to-end behavior of the proposed method more explicit, this subsection presents a concrete example for group128, one of the groups listed in Table 5. This group is particularly useful as an illustrative case because it exhibits a moderate novelty profile (30.39%), which allows the recommendation mechanism to be interpreted as a balance between thematic proximity and exploratory capacity, rather than as an extreme case of either very low or very high novelty. Figure 10 illustrates the complete recommendation process for group128, from group construction and evidence aggregation to novelty computation, topic candidate generation, and final recommendation.



**Figure 10.** Recommendation pipeline example for group128.

As shown in Figure 10, group128 is first modeled from co-authorship evidence and associated publications. Next, the novelty index is computed to characterize the group's openness to exploratory topics. BERTopic is then used to generate topic candidates from the group corpus, and the final recommendation is obtained by prioritizing directions that remain relevant while introducing a moderate degree of novelty.

#### Step 1: Group construction and evidence aggregation.

For group128, the pipeline identifies three collaborating authors and aggregates their shared and individual publication evidence into a group representation. In the generated group file, this group is associated with three co-authored articles and a larger pool of 36 publications when all member outputs are considered jointly. At the co-authored level, the extracted group descriptors are concise and include terms such as cemented cooling stage, newly discovered Holocene, owing large ice, volcanic clasts analysed, and capped chimborazo. Although this first descriptor set is short, the broader pool of keywords extracted from the members' complete publication histories reveals a much clearer thematic profile centered on volcanology, geodynamics, and volcanic hazards, with phrases such

as *cayambe volcanic complex*, chimborazo compound volcano, cotopaxi guagua pichincha, eruptions andesitic stratovolcano, and tephra fallout. This illustrates the intended behavior of the profiling stage: a compact group signature is first formed from shared work, and then enriched by the wider expertise of the members.

Step 2: Novelty index computation.

Following the formulation introduced in Section 3.8, novelty is derived from cosine similarity between the modeled group profile and the candidate topical space. For group128, the stored cosine similarity value is:

$$\text{sim}(\mathbf{P}_g, \mathbf{e}_t) = 0.6961 \quad (8)$$

and the novelty index is computed as:

$$\text{Nov}(g, t) = 1 - \text{sim}(\mathbf{P}_g, \mathbf{e}_t) \quad (9)$$

which yields:

$$\text{Nov}(\text{group128}) = 1 - 0.6961 = 0.3039 \quad (10)$$

or, equivalently, 30.39%. This value indicates that the group is not purely conservative, but neither is it highly exploratory. In practical terms, the recommendation mechanism should privilege topics that remain close to the group's volcanological expertise while still allowing some movement toward less central but still meaningful directions.

Step 3: Topic candidate generation.

Using the group-associated corpus, BERTopic produces a ranked set of 20 topic candidates. For group128, the highest-scoring candidate topics are dominated by volcanology- and geophysics-related descriptors. The first ranked topic is characterized by the terms:

eruption, volcano, magma, eruptions, eruptive, ash, volcanic, tungurahua, tephra, volcanoes

with a final score of 10.99. The second ranked topic is:

seismic, volcano, events, classification, volcanic, dcnn, tectonic, features, cotopaxi, feature

with a final score of 10.58. Additional high-ranked candidates include topics centered on arc, mantle, crustal, slab, subduction, rocks, earthquake, seismic, subduction, slip, and buildings, seismic, structures. These outputs show that the topic modeling stage does not generate arbitrary keyword bundles; rather, it recovers coherent candidate themes that remain strongly aligned with the geoscience profile of the group.

Step 4: Final recommendation.

After ranking the topic candidates, the final recommendation list stored for group128 includes the following leading topic descriptors:

rocks magmatic, eruption volcano magma, eruptions eruptive, volcano events, magma eruptions eruptive, classification volcanic, volcano events classification, seismic volcano events, volcanic subduction rocks, and eruptive ash volcanic

This list is fully consistent with the group profile reconstructed in Step 1. The recommendation is neither a generic geoscience output nor a trivial repetition of the original descriptors. Instead, it reorganizes the group's thematic evidence into a more actionable topic space that can support collective research planning. This worked example clarifies how the proposed pipeline operates from end to end: author-level evidence is first aggregated into a group profile through keywords and semantic representations, the similarity-based novelty formulation then quantifies how exploratory the recommendation can be, BERTopic trans-

forms the group corpus into interpretable topic candidates, and the ranking stage selects a recommendation set that remains domain-coherent while controlling the amount of novelty introduced. In the case of group128, the method behaves as expected for a moderately novel group, producing recommendations that stay close to the group's volcanological expertise while broadening the thematic space toward adjacent directions such as volcanic hazard classification, tectonic interpretation, and subduction-related processes.

## 5.2. Discussion and Limitations

The proposed pipeline demonstrated effective results in the creation of group-level topic recommendations by combining semantic profile modeling with advanced topic modeling techniques. However, there are multiple factors and nuances to consider regarding its performance, applicability, and areas for further improvement. This section addresses these aspects by analyzing the strengths and limitations of the approach, drawing comparisons with existing solutions, and providing insights into how the methodology can be optimized for real-world deployment.

The empirical results obtained from the keyword modeling and group recommendation stages show that the proposed system successfully balances precision and recall, ensuring that the generated topics are both relevant and novel for academic groups. Specifically, the integration of SentenceBERT embeddings with KeyBERT for keyword extraction offers a substantial improvement over traditional TF-IDF and CF-IDF models, which tend to either lose precision or lack semantic richness. The KeyBERT model, backed by the power of sentence embeddings, is able to generate contextually relevant keywords that align well with the underlying topics of the research corpus, thus improving the quality of recommendations.

Moreover, the use of a semantic space for author profile aggregation, as seen in the implementation of group embeddings, directly addresses the challenge of long-tail data. In typical academic datasets, where the majority of authors contribute only a small fraction of the publications, the approach minimizes the bias introduced by over-represented prolific authors, ensuring that the group recommendation process remains inclusive and balanced.

Despite the promising results, several limitations should be acknowledged in the current implementation of the pipeline. One practical limitation is that the present instantiation relies on Scopus as the primary source for bibliographic acquisition. Scopus provides rich and structured metadata that is well suited for profile modeling and collaboration analysis; however, the current study reflects the scope enabled by the available access and retrieval configuration.

As a result, the reported corpus should be understood as one operational view of the scholarly landscape rather than a complete representation of all possible sources. An important direction for future work is to extend the acquisition layer to additional bibliographic providers and open scholarly infrastructures, such as Crossref, OpenAlex, or institutional repositories, in order to improve coverage, reduce source dependence, and strengthen the generalizability of the recommendation pipeline across different academic settings.

Another limitation arises from the complexity and data dependence of topic modeling. In the present study, BERTopic is applied to a relatively large Scopus-derived corpus with sufficient document-level evidence, which supports the generation of coherent topic candidates. However, if the methodology is applied to much smaller datasets (e.g., fewer than 500 documents) or to records containing only very short text fields, topic instability may appear. In those cases, there may be insufficient textual evidence to characterize authors and groups, and the resulting topics may become fragmented, overly generic, or sensitive to clustering parameters. This limitation can be addressed by enriching document representations with titles, abstracts, keywords, and metadata; aggregating documents at the

author or group level; adjusting parameters such as minimum topic size; applying topic merging; or evaluating short-text topic modeling alternatives.

A related limitation is that the current evaluation does not include a direct comparison between BERTopic and classical topic modeling baselines such as Latent Dirichlet Allocation (LDA). Such a comparison would be useful to quantify the effect of replacing a probabilistic bag-of-words topic model with an embedding-based topic modeling approach in the proposed pipeline. Future work will therefore include a sensitivity analysis over corpus size, document length, and topic modeling parameters, together with a comparative benchmark involving BERTopic, LDA, NMF, and other neural or contextualized topic models to evaluate topic coherence, diversity, stability, interpretability, and downstream recommendation quality.

One of the standout features of the proposed approach is its emphasis on the trade-off between relevance, novelty, and diversity in group recommendations. As seen in the analysis of long-tail distribution in Figure 7, academic recommendation systems are prone to reinforcing popular topics, leaving little room for the exploration of new or emerging research areas. The novelty index integrated into the recommendation algorithm successfully addresses this issue, ensuring that the recommended topics for each group are not only relevant to the group's research focus but also introduce a degree of novelty, encouraging the exploration of less-studied topics.

However, it is essential to recognize that the novelty–diversity trade-off remains a complex challenge. In practice, academic groups may have varying levels of openness to novelty depending on their research context and the immediate needs of their collaborative work. The current implementation treats novelty and relevance as a weighted sum, but future iterations of the system could benefit from more advanced negotiation-based models, where group members' individual tolerance for novelty is modeled and incorporated into the recommendation process dynamically.

Another methodological aspect that may affect the generated groups is the choice of similarity measure used in seed-driven group formation. The current implementation uses cosine similarity because it is consistent with the Sentence–BERT embedding space and emphasizes semantic orientation rather than vector magnitude. Nevertheless, other similarity or distance measures could alter the local neighborhood of each seed author and, therefore, change the composition of seed-driven groups. Future work will include a sensitivity analysis comparing cosine similarity with Euclidean distance, Manhattan distance, and correlation-based similarity to quantify their impact on group formation and downstream topic recommendation.

Scalability remains one of the key concerns in deploying this recommendation system in a real-world academic setting. As highlighted in the Results section, the long-tail distribution of author productivity can significantly impact the performance of the system, especially when processing large datasets that span multiple research domains. The approach's ability to handle this challenge through incremental caching and coverage-based filtering is promising, but further optimization is needed.

For example, the current solution relies heavily on document-level embeddings and profile aggregation, which, although effective, can be computationally expensive as the dataset grows. Future work could explore methods for approximating these embeddings through techniques such as clustering or dimensionality reduction, or even integrating more lightweight models for larger-scale applications. Furthermore, the ability to scale the system for real-time recommendations—such as during collaborative research meetings or conferences—will require further advances in both algorithmic speed and infrastructure.

When compared to other scholarly recommender systems, our approach brings a unique advantage by combining group-specific profiling with advanced semantic models

like SentenceBERT and KeyBERT. Traditional recommender systems for academic contexts often rely on either citation-based approaches (e.g., collaborative filtering) or content-based approaches (e.g., TF-IDF, LDA) [15,41]. These systems are often limited by their inability to generate novel or diverse suggestions that are aligned with the actual collaborative dynamics of academic groups. Our work builds on these methods while introducing a more holistic model that accounts for group dynamics, group-level preferences, and novelty tolerance—filling a gap not adequately addressed in the existing literature.

In comparison with other group recommender systems [4,42], our system focuses specifically on the academic domain and incorporates sophisticated text representation methods that go beyond basic keyword matching. This makes our approach more adaptable to the complex, interdisciplinary, and rapidly evolving nature of academic research.

A further limitation concerns the scope of the profile modeling comparison. The experiments compare the proposed semantic keyword-based profiling component with TF-IDF and CF-IDF baselines, which are appropriate lexical and class-based descriptor baselines for the current evaluation. However, recent scientific representation models such as SPECTER and SciNCL provide stronger document-level alternatives that incorporate citation context and contrastive learning. We did not include these models as full profile modeling baselines in the present version because the pipeline was designed to operate over heterogeneous bibliographic metadata where complete citation graph information is not always available. Future work will evaluate alternative scientific encoders and researcher representation strategies as interchangeable profile modeling components within the same group recommendation framework.

## 6. Conclusions and Future Work

This study addressed the challenge of information overload and topic saturation in scholarly environments, where groups of researchers must identify relevant directions while avoiding repeated or overly mainstream themes. The problem becomes particularly pronounced under the long-tail dynamics of academic production, in which a small fraction of authors concentrate a large portion of publications, while many others contribute sparsely. In this context, designing a group-level recommendation mechanism requires (i) reliable and interpretable profile representations, (ii) mechanisms to balance relevance with novelty and diversity, and (iii) an evaluation strategy that reflects realistic collaboration settings.

The proposed approach contributes a unified pipeline that starts from bibliographic retrieval and preprocessing, then builds author and group profiles using semantic representations, and finally produces group-specific topic recommendations through a topic modeling stage. A key element of the solution is the profile modeling strategy, where author representations are derived from content-based signals and then aggregated to represent collaborative groups. Related work has shown that bibliographic traces and keyword-centered representations can support academic profile modeling [10]. In this study, these ideas are integrated with Sentence-BERT-based semantic embeddings, KeyBERT-style keyword extraction, and BERTopic-based topic modeling to produce recommendations that are not limited to lexical overlap but instead reflect contextual similarity and semantic cohesion.

The empirical results support three main outcomes. First, the distributional analysis confirms that scholarly corpora exhibit strong long-tail behavior, making naive frequency-based profiling prone to bias toward prolific authors; this motivates the use of semantic profiling and coverage-based perspectives. Second, the comparative evaluation indicates that embedding-driven keyword modeling (KeyBERT) improves the trade-off between precision and recall when compared with TF-IDF and CF-IDF, yielding more meaningful descriptors for downstream grouping and topic discovery. Third, the overall recommendation stage demonstrates that the pipeline can generate topic suggestions that remain

aligned with group interests while injecting novelty, reducing the tendency to repeatedly surface dominant themes in the dataset.

Despite these contributions, several limitations define the boundaries of the current implementation. The current implementation is instantiated on Scopus as the primary bibliographic source, which provides a structured and operationally consistent basis for profile modeling and recommendation. At the same time, this choice opens a clear opportunity to broaden the acquisition layer through additional scholarly sources and open infrastructures in order to expand coverage and strengthen the portability of the pipeline across different academic settings.

In addition, topic modeling remains sensitive to parameter choices (e.g., clustering granularity), and some generated topics may require refinement to improve interpretability. From an evaluation perspective, offline metrics and scenario-based analysis provide strong evidence of feasibility, yet user-centric studies with real research groups are required to quantify perceived usefulness, trust, and the practical value of novelty in collaborative decision-making.

Future work will focus on strengthening both the methodological and practical aspects of the system, with particular emphasis on temporal modeling and explainability:

- **Temporal and evolutionary profile modeling:** Extend the current static representation into a time-aware framework capable of capturing how author interests, collaboration patterns, and group themes evolve across years. This would allow the recommender to distinguish between long-term expertise and emerging directions, making topic suggestions more sensitive to recent trajectories and shifts in collaborative research agendas.
- **Explainability and transparent recommendation rationales:** Incorporate explicit explanation mechanisms so that each recommended topic can be traced back to the profile evidence that supports it, such as representative keywords, supporting publications, similarity relations, and novelty signals. This would improve transparency, trust, and practical usability, especially in academic settings where recommendations are expected to be discussed, justified, and validated collectively.
- **Comparative topic modeling benchmarks:** Extend the evaluation by comparing BERTopic with classical and alternative topic modeling approaches, including LDA, NMF, and contextualized neural topic models. This comparison should consider topic coherence, topic diversity, topic stability, interpretability, and the effect of each topic model on downstream group recommendation quality.
- **Multi-source scholarly ingestion:** Extend retrieval beyond Scopus to incorporate open scholarly graphs, Crossref, OpenAlex, and institutional repositories, improving coverage and reducing dependence on a single bibliographic provider.
- **Preference- and feedback-aware recommendation:** Introduce lightweight feedback loops where groups can validate, refine, or reject suggested topics, allowing the system to learn group-level novelty tolerance and recommendation preferences dynamically.
- **Scalability and deployment:** Optimize embedding computation and retrieval through vector indexing, incremental updates, and more efficient large-scale processing strategies, enabling near real-time recommendations for institutional-scale corpora.
- **User studies and comparative evaluation protocols:** Conduct controlled studies with academic groups to evaluate perceived relevance, explanation quality, novelty acceptance, and decision impact, complementing offline metrics with human-centered evidence.

**Author Contributions:** Conceptualization, C.A.-T., L.R. and E.L.-A.; methodology, L.R. and E.L.-A.; software, C.A.-T.; validation, L.R. and E.L.-A.; formal analysis, C.A.-T. and L.R.; investigation, C.A.-T.,

L.R. and E.L.-A.; resources, C.A.-T.; data curation, C.A.-T.; writing—original draft preparation, C.A.-T.; writing—review and editing, C.A.-T., L.R. and E.L.-A.; supervision, L.R. and E.L.-A.; project administration, L.R. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by Escuela Politécnica Nacional.

**Data Availability Statement:** The raw data supporting the conclusions of this article will be made available by the authors on request.

**Acknowledgments:** The authors would like to thank Faculty of Systems Engineering, “Departamento de Informática y Ciencias de la Computación”, Escuela Politécnica Nacional, Quito 170517, Ecuador for academic and financial support.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Larsen, P.O.; von Ins, M. The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index. *Scientometrics* **2010**, *84*, 575–603. [CrossRef] [PubMed]
2. Fortunato, S.; Bergstrom, C.T.; Börner, K.; Evans, J.A.; Helbing, D.; Milojević, S.; Petersen, A.M.; Radicchi, F.; Sinatra, R.; Uzzi, B.; et al. Science of science. *Science* **2018**, *359*, eaao0185. [CrossRef] [PubMed]
3. Ricci, F.; Rokach, L.; Shapira, B. Recommender systems: Techniques, applications, and challenges. In *Recommender Systems Handbook*; Ricci, F., Rokach, L., Shapira, B., Eds.; Springer: New York, NY, USA, 2022; pp. 1–35. [CrossRef]
4. Felfernig, A.; Boratto, L.; Stettinger, M.; Tkalčič, M. *Group Recommender Systems: An Introduction*; Springer: Berlin/Heidelberg, Germany, 2018. [CrossRef]
5. Masthoff, J. Group Recommender Systems: Aggregation, Satisfaction, and Group Attributes. In *Recommender Systems Handbook*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 743–776. [CrossRef]
6. McNee, S.; Riedl, J.; Konstan, J. Being accurate is not enough: How accuracy metrics have hurt recommender systems. In *Proceedings of the CHI'06 Extended Abstracts on Human Factors in Computing Systems*; Association for Computing Machinery: New York, NY, USA, 2006; pp. 1097–1101. [CrossRef]
7. Castells, P.; Hurley, N.J.; Vargas, S. Novelty and Diversity in Recommender Systems. In *Recommender Systems Handbook*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 881–918. [CrossRef]
8. Visser, M.; van Eck, N.J.; Waltman, L. Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic. *Quant. Sci. Stud.* **2021**, *2*, 20–41. [CrossRef]
9. Elsevier. Elsevier Scopus APIs, 2023. Available online: [https://dev.elsevier.com/guides/Scopus%20API%20Guide\\_V1\\_20230907.pdf](https://dev.elsevier.com/guides/Scopus%20API%20Guide_V1_20230907.pdf) (accessed on 12 April 2025).
10. Ayala-Tipan, C.; Recalde, L.; Suntaxi, G. Enhancing Academic Profiling with Advanced NLP Techniques: SBERT + KeyBERT. In *Proceedings of the Information Management and Big Data*; Lossio-Ventura, J.A., Ceh-Varela, E., Díaz, E., Paz Espinoza, F., Tadonki, C., Alatrasta-Salas, H., Eds.; Springer: Cham, Switzerland, 2025; pp. 195–211. [CrossRef]
11. Beel, J.; Gipp, B.; Langer, S.; Breiteringer, C. Research-paper recommender systems: A literature survey. *Int. J. Digit. Libr.* **2016**, *17*, 305–338. [CrossRef]
12. Kreutz, C.K.; Schenkel, R. Scientific paper recommendation systems: A literature review of recent publications. *Int. J. Digit. Libr.* **2022**, *23*, 335–369. [CrossRef] [PubMed]
13. Park, Y.J.; Tuzhilin, A. The Long Tail of Recommender Systems and How to Leverage It. In *Proceedings of the 2008 ACM Conference on Recommender Systems (RecSys '08)*; ACM: New York, NY, USA, 2008; pp. 11–18. [CrossRef]
14. Abdollahpouri, H.; Burke, R.; Mobasher, B. Controlling Popularity Bias in Learning-to-Rank Recommendation. In *Proceedings of the Eleventh ACM Conference on Recommender Systems (RecSys '17)*; ACM: New York, NY, USA, 2017; pp. 42–46. [CrossRef]
15. Kaminskas, M.; Bridge, D. Diversity, Serendipity, Novelty, and Coverage: A Survey and Empirical Analysis of Beyond-Accuracy Objectives in Recommender Systems. In *ACM Transactions on Interactive Intelligent Systems*; Association for Computing Machinery: New York, NY, USA, 2017; Volume 7. [CrossRef]
16. Abdollahpouri, H.; Burke, R.; Mobasher, B. User-centered Evaluation of Popularity Bias in Recommender Systems. In *Proceedings of the ACM on Human-Computer Interaction*; Association for Computing Machinery: New York, NY, USA, 2021; Volume 5. [CrossRef]
17. Burke, R. Hybrid Recommender Systems: Survey and Experiments. *User Model.-User Interact.* **2002**, *12*, 331–370. [CrossRef]
18. Adomavicius, G.; Tuzhilin, A. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Trans. Knowl. Data Eng.* **2005**, *17*, 734–749. [CrossRef]

19. Cohan, A.; Feldman, S.; Beltagy, I.; Downey, D.; Weld, D.S. SPECTER: Document-level Representation Learning using Citation-informed Transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 2270–2282. [\[CrossRef\]](#)
20. Beel, J.; Aizawa, A.; Breiteringer, C.; Gipp, B. Mr. DLib: Recommendations-as-a-Service (RaaS) for Academia. In *Proceedings of the 2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*; IEEE: New York, NY, USA, 2017; pp. 1–2. [\[CrossRef\]](#)
21. Ekstrand, M.; Harper, F.; Willemsen, M.; Konstan, J. User perception of differences in recommender algorithms. In *Proceedings of the 8th ACM Conference on Recommender Systems*; Association for Computing Machinery: New York, NY, USA, 2014; Volume 10. [\[CrossRef\]](#)
22. Zhang, Z.; Patra, B.G.; Yaseen, A.; Zhu, J.; Sabharwal, R.; Roberts, K.; Cao, T.; Wu, H. Scholarly recommendation systems: A literature survey. *Knowl. Inf. Syst.* **2023**, *65*, 4433–4478. [\[CrossRef\]](#)
23. Shiddiqi, A.M.; Alzamzami, M.N.; Adillion, I.G.; Budiman, M.I.A.; Supriyanto, R.; Machmud, M. Findme-Scholar: A contextual researcher recommender system for enhancing research collaboration using adaptive topic interest area modelling. *MethodsX* **2025**, *14*, 103583. [\[CrossRef\]](#) [\[PubMed\]](#)
24. Pérez-Almaguer, Y.; Yera, R.; Alzahrani, A.A.; Martínez, L. Content-based group recommender systems: A general taxonomy and further improvements. *Expert Syst. Appl.* **2021**, *184*, 115444. [\[CrossRef\]](#)
25. Sacharidis, D. Top-N group recommendations with fairness. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*; SAC '19; Association for Computing Machinery: New York, NY, USA, 2019; pp. 1663–1670. [\[CrossRef\]](#)
26. Amatriain, X.; Jaimes, A.; Oliver, N.; Pujol, J. *Data Mining Methods for Recommender Systems*. In *Recommender Systems Handbook*; Springer: Boston, MA, USA, 2011; pp. 39–71. [\[CrossRef\]](#)
27. de Campos, L.M.; Fernández-Luna, J.M.; Huete, J.F.; Ribadas-Pena, F.J.; Bolaños, N. Information Retrieval and Machine Learning Methods for Academic Expert Finding. *Algorithms* **2024**, *17*, 51. [\[CrossRef\]](#)
28. Ostendorff, M.; Rethmeier, N.; Augenstein, I.; Gipp, B.; Rehm, G. Neighborhood Contrastive Learning for Scientific Document Representations with Citation Embeddings. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*; Association for Computational Linguistics: New York, NY, USA, 2022; pp. 11670–11688. [\[CrossRef\]](#)
29. Nagao, H.; Katsurai, M. Researcher Representations Based on Aggregating Embeddings of Publication Titles: A Case Study in a Japanese Academic Database. In *Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024)*, Bangkok, Thailand, 15–16 August 2024; pp. 277–282.
30. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 4171–4186. [\[CrossRef\]](#)
31. Reimers, N.; Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 3982–3992. [\[CrossRef\]](#)
32. Newman, M.E.J. The structure of scientific collaboration networks. *Proc. Natl. Acad. Sci. USA* **2001**, *98*, 404–409. [\[CrossRef\]](#) [\[PubMed\]](#)
33. Elsevier. Scopus APIs. 2023. Available online: <https://dev.elsevier.com/scopus.html> (accessed on 18 December 2025).
34. Sarwar, T.B.; Noor, N.M.; Miah, M.S.U. Evaluating keyphrase extraction algorithms for finding similar news articles using lexical similarity calculation and semantic relatedness measurement by word embedding. *PeerJ Comput. Sci.* **2022**, *8*, e1024. [\[CrossRef\]](#) [\[PubMed\]](#)
35. Bianchi, F.; Terragni, S.; Hovy, D. Pre-training is a Hot Topic: Contextualized Document Embeddings Improve Topic Coherence. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 759–766. [\[CrossRef\]](#)
36. McInnes, L.; Healy, J.; Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *J. Open Source Softw.* **2018**, *3*, 861. [\[CrossRef\]](#)
37. McInnes, L.; Healy, J.; Astels, S. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.* **2017**, *2*, 205. [\[CrossRef\]](#)
38. Yan, X.; Guo, J.; Lan, Y.; Cheng, X. A Biterm Topic Model for Short Texts. In *Proceedings of the 22nd International Conference on World Wide Web*; Association for Computing Machinery: New York, NY, USA, 2013; pp. 1445–1456. [\[CrossRef\]](#)
39. Johnson, J.; Douze, M.; Jégou, H. Billion-scale similarity search with GPUs. *IEEE Trans. Big Data* **2019**, *7*, 535–547. [\[CrossRef\]](#)
40. Röder, M.; Both, A.; Hinneburg, A. Exploring the Space of Topic Coherence Measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining (WSDM '15)*; Association for Computing Machinery: New York, NY, USA, 2015; pp. 399–408. [\[CrossRef\]](#)

41. Bobadilla, J.; Ortega, F.; Hernando, A.; Gutiérrez, A. Recommender systems survey. *Knowl.-Based Syst.* **2013**, *46*, 109–132. [[CrossRef](#)]
42. Boratto, L.; Carta, S.; Chessa, A.; Agelli, M.; Clemente, M.L. Group Recommendation with Automatic Identification of Users Communities. In *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*; IEEE: New York, NY, USA, 2009; Volume 3, pp. 547–550. [[CrossRef](#)]

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.