Available online at website: https://jurnal.iaii.or.id/index.php/RESTI

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 9 No. 3 (2025) 497 - 505 e-ISSN: 2580-0760

Expertise Retrieval Using Adjusted TF-IDF and Keyword Mapping to ACM Classification Terms

Lyla Ruslana Aini[1*], Evi Yulianti[1]
[1,2]Department of Computer Science, Faculty of Computer Science, Universitas Indonesia, Depok, Indonesia
[1]Research Center for Data and Information Sciences (PRSDI), National Research and Information Agency (BRIN), KST Samaun Samadikun Bandung, Indonesia

[1]lyla.ruslana@ui.ac.id, lyla001@brin.go.id, [2]evi.y@cs.ui.ac.id

Abstract

In an era of collaboration, knowing someone's expertise is becoming increasingly necessary. Recognizing individuals' proficiency can be challenging because it requires considerable manual time. This study explores the expertise of lecturers from the Computer Science Department, Universitas Indonesia (Fasilkom UI), based on scientific publications. The data were obtained from the Sinta journal website's scrapping process, which includes Scopus, Garuda, and Google Scholar data sources. The approach used was keyword extraction using the adjusted TF-IDF. The resulting keywords were then mapped to the ACM classification class using cosine similarity calculations with various embedding models, including BERT, BERT multilingual, FastText, XLM Roberta, and SBERT. The experimental results highlighted that combining the adjusted TF-IDF with mapping to the ACM classes using SBERT is a promising approach for gaining the best expertise. The use of abstract data has proved to be better than that of full-text data. Using title-abstract-EN data achieved a score of 0.49 for both the P@1 and NDCG@1 metrics, whereas the title-abstract-ENID data attained a score of 0.75 for both metrics P@1 and NDCG@1.

Keywords: adjusted TF-IDF; ACM classification; BERT; expertise; FastText; BERT multilingual; SBERT; XLM-RoBERTA

How to Cite: L. R. Aini and Evi Yulianti, "Expertise Retrieval Using Adjusted TF-IDF and Keyword Mapping to ACM Classification Terms", *J. RESTI (Rekayasa Sist. Teknol. Inf.)* , vol. 9, no. 3, pp. 497 - 505, May 2025.
Permalink/DOI: https://doi.org/10.29207/resti.v9i3.6397

Received: February 16, 2025
Accepted: May 5, 2025
Available Online: May 25, 2025

1. Introduction

In the last twelve years, there has been significant interest and a vast array of findings in expertise retrieval, an evolving subfield of information retrieval (IR) [1]. This field has gained significant importance in various domains, including job seeker profiling, expert finding, and people or institution collaboration [2], [3] assisting in the recruitment process for job candidates [4]. An expert is a person who completes specific tasks [5]. Expertise depends on knowledge (what an expert knows) and skill (what an expert knows how to do). The skills examined are hard skills, which involve specialized knowledge and focus on solving concrete problems [6]. An essential element of expertise is the intuition to recognize several relevant states that arise in any given situation and then to retrieve information from memory about what to do when those states arise [7]. Expertise falls into the category of tacit knowledge, which is challenging to express, acquired through experience, self-learning, and influenced by beliefs, perspectives, and values. The way to obtain expertise from tacit knowledge is by extracting evidence of expertise, for example, written documents, electronic communications, and social networks [2]. Expertise retrieval systems aim to connect users with the most relevant experts by analyzing large corpora of text, such as publications, patents, and online content, and leverage NLP techniques to analyze corpus data to determine the expertise and influence of individuals. The study of [8] states that the expertise of researchers is characterized based on the distribution of topics in the researcher's papers.

Many of the initial automated expertise retrieval systems often concentrated on particular types of documents. Study [9] attempted to identify expertise within email communications, as emails naturally reflect potential experts' activities, interests, and objectives. [10] suggest multiple approaches to

automatically define areas of expertise by extracting keywords from publications employing term frequency-inverse document frequency (TF-IDF), incorporating title weighting, and applying a keyword merging method. Citation-based approach was introduced by [11] with Context-based cluster analysis (CCA) and Cluster-based Ontology Generation framework (COGA). Hybrid models [12] are also raised to estimate the level of expertise (weight) of an expert in a topic by weighting IDF with n-grams, GM (Graph-based Model), and ECG (expert-collaboration graph), which represents the relationship between experts and documents based on co-authoring information.

Another approach explores collaborations to identify the strengths of an institution's core or individual potential expertise[3], [13] and competencies and seek interaction between developers [14]-[17]. Thematic area strengths are obtained by looking at citations from publications and the number of publications. The author profiling at the community level is carried out by [18], [19]. Authors are linked by topic distribution and community distribution [18]. Moreover, many skills are extracted from the QA community[19]-[22].

We investigate the expertise of lecturers in the Computer Science Department at Universitas Indonesia (Fasilkom UI) through their scientific publications. The method employed involves keyword extraction using the adjusted TF-IDF approach with title weighting, as done by [10]. The results are then mapped to the ACM (Association for Computing Machinery) classification terms based on cosine similarity score utilizing various embeddings models such as BERT, BERT Multilingual (MBERT), FastText, XLM Roberta, and Sentence Transformers (SBERT). Our main contributions in this study include:

Integration of keyword extraction with the adjusted TDF-IF baseline method with classes in the ACM classification. Previous studies [10] extracted keywords but did not systematically map them to a widely accepted taxonomy. We address this by mapping the adjusted TF-IDF results to the ACM CCS categories to improve interpretability and relevance to expertise terms.

We explore word embeddings (FastText, BERT, MBERT, SBERT, and XLM-RoBERTA) to determine the best embedding model for mapping keywords to ACM categories, which are rarely systematically compared in expertise search studies.

Comparing different data resources, namely abstract data with full-text data and English datasets with mixed-language datasets (Indonesian-English), provides insights into which type of document structure produces the best expertise search results.

2. Methods

The general IR system in this study is shown in Figure 1, where the input is a lecturer's name query with lecturer publication documents as a data source for the retrieval process, and the output is top-k lecturer expertise.

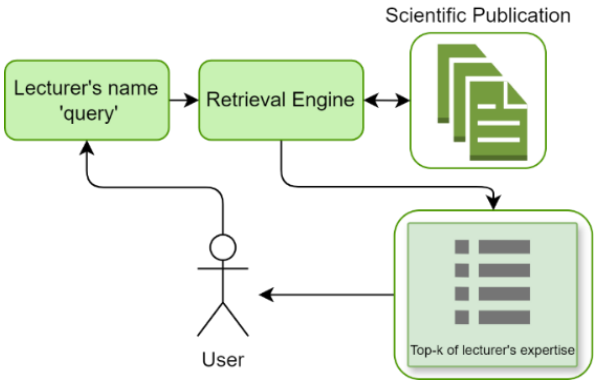

Figure 1. General IR System for Expertise Retrieval

Figure 2 depicts the methodology utilized in this research, which comprises three primary processes. The first process pertains to data scrapping from the Sinta website (sinta.kemdikbud.go.id). The second process involves keyword extraction from the data collected. Moreover, the third process is mapping expertise terms to the ACM class, which involves various word or phrase embedding based on their cosine distance values.

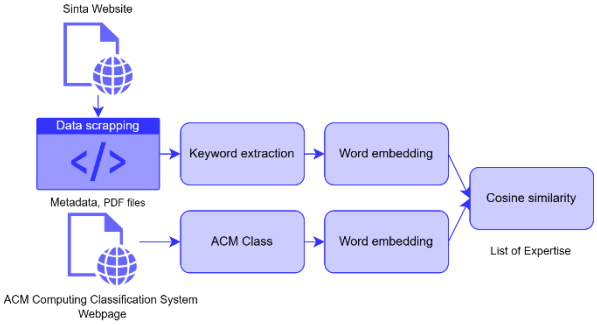

Figure 2. Research Methodology for Expertise Retrieval

2.1 Data Scrapers for Scientific Publication Document

We created web scrapers using Python's Beautiful Soup libraries to extract data from HTML and utilized urllib to gather information about the HTML page. A session must be implemented due to the restriction of viewing the Sinta website without login. The scrapping process was performed based on the lecturer's name at the Computer Science Department Universitas Indonesia (Fasilkom UI) and the affiliate 'Universitas Indonesia'. After obtaining the appropriate author, the provided document data is scrapped. Sinta provides five academic document sources: Scopus, Garuda, Google Scholar, Web of Science, and Rama. This research collects data from Scopus, Garuda, and Google Scholar. The Scopus and Google Scholar page scrapping uses an API provided through the registration process on the Elsevier (https://dev.elsevier.com/) and Google developer page (https://serpapi.com/google-scholar-api). Meanwhile, the scrapping process is carried out for Garuda pages using a script from scratch. Figure 3

shows the stages of data collection and processing from the Sinta journal.

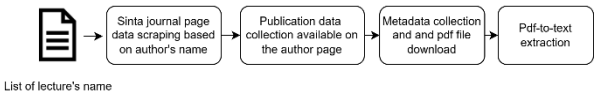

Figure 3. Data Collection and Preprocessing

Metadata information obtained from the Scopus page includes names of all authors and author IDs, affiliation of each author, publication title, publication date, abstract, keywords, data source address link, and doi. Garuda page metadata has additional data taken, namely RIS and Bibtex citation information, Google Scholar addresses, and Full-Text PDF addresses. PDFs are obtained only from open-access pages on both the Scopus and Garuda pages. The total number of Fasilkom UI lecturer authors acquired was forty-five authors. The data obtained for this study are listed in Tables 1 and 2. Most of the published data obtained were in English. Figure 4 shows a sample snippet of the metadata extracted from the scraping process.

Table 1. Data Collection from Sinta Website

Source	Abstract	Full-Text PDF
Scopus	1811	285
Garuda	145	200
Google Scholar	-	13

Table 2. Language Comparison in the Data Collection

Document	English	Indonesia
Abstract	1881	28
Full-text	301	72

Code Generator Development to Transform IFML (Interaction Flow Modelling Language) into a React-based User Interface⇥Model-Driven Software Engineering (MDSE) is a software development approach that uses the Model to be the main actor of the development. MDSE can be applied to User Interface (UI) Development so that a model for the UI can be built, and then a transformation can be made to turn it into a running application. In this research, we develop UI Generator to support UI Development with the MDSE approach. This UI Generator can also support UI Development in Software Product Line Engineering (SPLE) paradigm. The UI is modeled with Interaction Flow Modeling Language (IFML) diagram. Then The IFML diagram is transformed into React-Based UI by the UI Generator. The UI Generator is developed with Acceleo on Eclipse IDE to transform IFML into React Code with the transformation rules defined in this research. The UI generator is also enriched with display settings and static page management to address user customization needs. The experimental results show that the UI Generator can generate a functional website. Besides evaluating the working product, UI Generator is evaluated qualitatively well based on six quality criteria as an SPLE supporting tool.⇥Ilma Ainur Rohma|Universitas Indonesia|/author/view/7774984###Ade Azurat|Unknown|/author/view/507155⇥Jurnal Ilmu Komputer dan Informasi⇥Vol. 17 No. 2 (2024): Jurnal Ilmu Komputer dan Informasi (Journal of Computer Science and Informatio⇥02 Jun 2024⇥/citation/site/RIS/4563070⇥/citation/site/bib/4563070⇥ https://jiki.cs.ui.ac.id/index.php/jiki/article/view/1178/509⇥ https://jiki.cs.ui.ac.id/index.php/jiki/article/download/1178/509⇥http://scholar.google.com/scholar?q=%2Bintitle%3A"Code+Generator+Development+to+Transform+IFML+Interaction+Flow+Modelling+Language+into+a+React-based+User+Interface"⇥https://garuda.kemdikbud.go.id/documents/detail/4563070⇥ade+azurat+6025983

(4a) Metadata collected from the Garuda web

https://www.scopus.com/record/display.uri?eid=2-s2.0-85191336941&origin=resultslist⇥01-01-2023⇥ Indonesia###60122051###None###Universitas Trunojoyo Madura###Department of Informatics⇥A Novel Deterministic Algorithm for Optimizing Workflow Discovery with Short-Loops in Large Event-Log⇥© 2023 IEEE.One of the main problems in process discovery is obtaining correctness workflow when analysing a Large Event-Log with high concurrent control flow, complex short loops, and uncertainty. T+ is a novel method to construct a standardization workflow model in the SWF-Net for discovering a large event log with high concurrency free-choice, parallel, and non-free-choice control flow within an uncertainty short-loop. By the testing result, the short-loop construction places with the optimistic and pessimistic One-Loop-Free approach have the best results compared to other existing methods based on deterministic, heuristic, genetic, inductive, also Integer Linear Programming algorithms. T+ has proofed achieved completeness and correctness SWF-Net with high rate on fitness recall up to 95% and appropriateness up to 90%.⇥ http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=10490443⇥ 10.1109/ICTECA60133.2023.10490829⇥component###formatting###insert###style###styling

(4b) Metadata collected from the Scopus web

Figure 4. A Snippet of Metadata Extracted from the Scraping Process

2.2 Keyword Extraction

A specialist's expertise can be represented by words drawn from their papers, abstracts, and titles or by the keywords linked to their documents [2]. One method used is keyword extraction. Keyword (also known as a key phrase or key term) is typically defined as a sequence of one or more words that concisely represent the content of a document [23]. Keyword extraction identifies the lexical units most effectively represent the document [24]. Keyword extraction involves automatically extracting significant and characteristic words or phrases from a document to express its key content aspects. Two approaches are commonly used: unsupervised and supervised [25], [26], [27]. Term Frequency is a statistically based unsupervised approach to deriving words or phrases from documents [28]. Study [8] constructs expertise with the TF-IDF approach, title weighting, and a keyword merging technique. This study uses the TF-IDF method and maps the result to ACM classification to better expertise representation.

2.3 Adjusted TF-IDF

The adjusted TF-IDF approach [10] improves upon the traditional TF-IDF by taking the word/phrase in the publication keyword section as the words calculated in the TF-IDF calculation and combining it with title weighting, which results in better expertise extraction for R&D publication data. The method helps prioritize the keywords created by the authors that represent the publication content over other terms appearing in documents. This adjusted TF-IDF also considers the importance of the keywords appearing in the publication title by giving them higher weights than

those found in other sections. This weighting aligns with the idea that the title is a summary that concisely describes the research focus and is relevant to the document's content. The evaluation results show that the modified TF-IDF approach provides better expertise results by the increase in F-measure from 24.6% to 31.6% [10]. The method provides meaningful keyword ranking improvements, making it more effective for building researcher expertise profiles from R&D publications. Based on these considerations, we use this method as a baseline in our study.

We collect keywords from the publication of each lecturer in list K. Then, TF-IDF is calculated based on the list of K. TF-IDF serves as the standard baseline for keyword extraction, evaluating, and prioritizing words/phrases based on a specific formula such as Equation 1.

$$AdjTFIDF(k) = \sum_{i=1}^{N} tf(k, d_i) \times \log \frac{D}{df(k)} \quad (1)$$

AdjustedTFIDF(*k*) represents the TFIDF score for each keyword *k* in list *K*, *tf*(*k*, d_i) denotes the frequency of keyword *k* in document *i*, *df*(*k*) refers to the number of documents containing keyword *k*, *D* stands for the total number of all documents, *N* is the total number of documents associated with each lecturer. A study of [1] considers keywords that appeared in the title of the publication as essential terms and, as a result, introduces the weighting scheme presented in Equation 2.

$$Score(k) = AdjTFIDF(k) \times \left(1 + \frac{df_{\prime}(k)}{\max\{df'(k)\}}\right) \quad (2)$$

Score (*k*) represents the score for each keyword in the list *K*, *df*'(*k*) indicates the number of document titles that include keyword k, *max*{*df*'(*k*)} is the highest value of *df*'(*k*).

This study undergoes two scenarios: extracting the keywords of the title and abstract and extracting the keywords of the entire contents of the scientific publication document. Some preprocesses on text data are the removal of punctuation, apostrophes, stop-words, and numbers.

2.3 ACM Classification

The keywords resulting from the adjusted TF-IDF are sometimes still technical and must reflect expertise terms. Therefore, it is necessary to map the term to the standard expertise term. This study uses the ACM classification standard. The terms in the ACM class will be the expertise terms that will result from this research. The ACM CCS classification system has become the de facto standard for classifying computing literature since 1964 [23]. Professor Zvi Kedem leads the CCS update project from NYU, along with 120 specialists in the field of computing. The latest version (2012) underwent two review processes and many iterations. The first draft of the ACM classification used data, including logs of user searches in the ACM digital library, analysis of author-provided text analysis and keyword occurrences, and review of existing computer science taxonomy guides. ACM domain experts use the draft to update CCS. CCS 2012 [24]was developed as a semantic ontology and is available in SKOS format. The summary of the ACM classification class number is described in Table 3.

Table 3. Class Number in ACM Classification Tree

Level	Class Number	Class Example
1st	13	General and reference, Mathematics of computing, Information systems, Security and privacy
2nd	84	Document types, Printed circuit boards, Architectures, Network architectures, Cryptography
3rd	553	Surveys and overviews, Serial architectures, Network design principles, Graph theory
4th	983	Machine translation, Operating systems, Software infrastructure, Routing protocols
5ft	333	Quantum communication and cryptography, Biometrics, Topic modeling, Virtual machines
6th	29	Embedded middleware, Randomized local search, Fiber distributed data interface (FDDI), CS1

2.3 Embedding Model

Word embeddings are a fundamental concept in natural language processing (NLP) [20] that transforms words into continuous vector representations [2][3]. These vectors capture semantic meanings and relationships between words, enabling machines to understand and process human language more effectively. Traditional methods, such as one-hot encoding, represent words as sparse vectors that do not capture the relationships between words. In contrast, word embeddings create dense, low-dimensional vectors that embed semantic information, making them powerful tools for various NLP tasks. This study elevates some embedding techniques, including BERT, MBERT, FastText, XLM Roberta, and SBERT.

BERT [21] is a transformer-based model designed for natural language understanding by pretraining on vast text data in a bidirectional manner [4]. BERT captures context from left and right surrounding words, making it highly effective for question-answering and sentence classification tasks. Multilingual BERT (MBERT) extends this capability by being pre-trained on text from over one hundred languages, enabling cross-lingual generalization. Despite not being explicitly trained for translation, MBERT competes strongly in zero-shot cross-lingual tasks, making it a versatile tool for multilingual natural language processing applications while SBERT enables efficient embedding of sentences [5]. BERT and MBERT are transformer-based models known for capturing contextual meaning and are helpful in mapping extracted keywords to expertise terms.

XLM-RoBERTa (XLM-R) is an extension of the RoBERTa model [22], optimized explicitly for cross-lingual tasks. Pretrained on over one hundred languages using masked language modeling, XLM-R achieves state-of-the-art performance in various multilingual benchmarks without relying on explicit parallel data. Its

robust architecture allows for effective transfer learning across languages, significantly improving results in both high-resource and low-resource languages. XLM-RoBERTA supports cross-lingual understanding, making it valuable for mixed-language datasets (Indonesian-English).

Sentence Transformers [23] build on BERT and its variants (including RoBERTa) by fine-tuning models to generate semantically meaningful sentence embeddings. Sentence Transformers allow for computationally efficient, high-quality sentence representations by applying techniques like contrastive learning and Siamese networks. SBERT is designed for sentence similarity tasks, making it suitable for cosine similarity-based mapping between keyword extraction results and ACM classification classes.

FastText [24] operates on the principles of Word2Vec and the n-grams technique. In Word2Vec, text is fed into the neural network individually. However, in FastText, words are divided into several subwords before being fed into the neural network. After training the neural network on the training data, a word vector is obtained for each n-gram. These n-grams can later be used to relate to other words, allowing for the mapping of rare words due to the overlapping n-grams in other words. We also considered the FastText model because we wanted to see whether context would affect the mapping of words extracted from adjusted TF-IDF keywords to ACM classification. The FastText model is selected based on the subword-based embedding model to handle OOV.

The pre-trained embedding models used in this study for BERT, MBERT [21], SBERT, XLM-RoBerta, and FastText models are bert-large-uncased, bert-base-multilingual-uncased, all-mpnet-base- v2, xlm-roberta-base [22] and cc.en.300.bin respectively. The embedding model converts the top ten keyword results from the adjusted TF-IDF and ACM classification terms into vectors (embeddings), which then calculate the cosine similarity between the two sets of terms as shown in Figure 5. The term expertise result is the distance with the highest cosine similarity value.

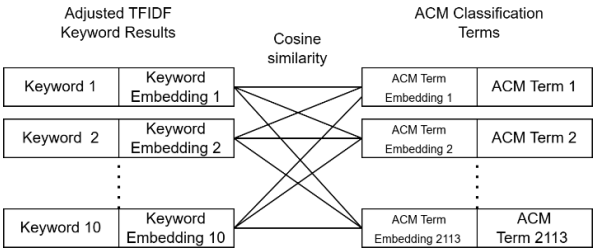

Figure 5. Keyword Mapping to ACM Classification Class

3. Results and Discussions

The expertise retrieval process involves two scenarios of data: the title-abstract and full-text data of the scientific publication document and in three types of language: English (EN), Indonesian (ID), and a mix of English and Indonesian (ENID). The number of lecturer data evaluated for EN data is forty-five, while for ID and ENID data, it is twelve because only twelve lecturers were collected with the two types of ID and EN publication data. In each scenario, keyword extraction is carried out with adjusted TF-IDF. The top ten keywords with the highest adjusted TF-IDF scores will be represented as vectors using embedding techniques, including BERT, MBERT, XLM-RoBERTA, SBERT, and FastText.

The cosine similarity between the vector of the adjusted TF-IDF keywords results and the ACM classification terms is calculated and then evaluated by three annotators. The annotator is giving relevance values of 1 and 0. A value of 1 is given if the resulting expertise matches the lecturer's expertise. Another term annotation was done to determine whether a term in the ACM classification is classified as an expertise term, with a value of 1 indicating an expertise term and zero if the term is considered not an expertise. Fleiss kappa values from all experimental scenarios show a range of 0.57~0.63, with little difference between the methods used. The evaluation metrics used are precision@k, MRR, and NDCG@k. Table 4 shows the evaluation results for English Data. Adjusted TF-IDF was used as the baseline method in this study.

Table 4. Evaluation Results for English Data.

Data	Method	P@3	P@1	NDCG@3	NDCG@1	MRR
Title-abstract	AdjTFIDF	0.4	0.42	0.4	0.42	0.61
	AdjTFIDF +SBERT	0.33	0.33	0.33	0.33	0.55
	AdjTFIDF +MBERT	0.06	0.11	0.07	0.11	0.18
	AdjTFIDF+BERT	0.01	0.04	0.02	0.04	0.09
	AdjTFIDF+FastText	0.1	0.13	0.1	0.13	0.23
	AdjTFIDF+XLM-R	0.01	0.02	0.01	0.02	0.06
Full-text	AdjTFIDF	0.3	0.27	0.29	0.27	0.45
	AdjTFID+SBERT	0.3	0.29	0.3	0.29	0.49
	AdjTFIDF+MBERT	0.11	0.11	0.11	0.11	0.29
	AdjTFIDF+BERT	0.12	0.11	0.12	0.11	0.25
	AdjTFIDF+FastText	0.28	0.33	0.29	0.33	0.5
	AdjTFIDF+XLM-R	0.09	0.11	0.1	0.11	0.24

The baseline method with adjusted TF-IDF (AdjTFIDF) shows the highest performance for title-abstract data with scores 0.4, 0.42, 0.4, 0.42, and 0.61, followed by AdjTFIDF+SBERT with scores 0.33, 0.33,

0.33, 0.33, and 0.55 for P@3, P@1, NDCG@3, NDCG@1, and MRR, respectively. The AdjTFIDF+FastText method results provide the best performance for full-text data for P@1, NDCG@1, and MRR metrics with scores 0.33, 0.33, and 0.5, followed by AdjTFIDF+SBERT for P@3 and NDCG@3 metrics with scores 0.3 and 0.3. The evaluation results also show that abstract data has a higher AdjTFIDF value than full-text data. Abstracts contain more specific words and are close to or identical to the terms in the publication keywords. Those publication keywords are used as predefined terms calculated using the AdjTF-IDF method and become a factor in increasing the AdjTFIDF score. Table 5 shows the expertise results sample of Lecturer-1 for English title-abstract and full-text data.

Based on the English data results, we apply the two best keyword mapping methods, AdjTFIDF+SBERT and AdjTFIDF+FastText, to Indonesian (ID) and a mix of English-Indonesian (EN-ID) data. If the keyword extraction result of AdjTFIDF is an Indonesian term, it will be translated into English using the 'langdetect' library before mapping to the ACM class. Table 6 shows the evaluation result of the ID and EN-ID data. The minimum number of resulting keywords for these data scenarios is three, so the evaluation calculation is carried out up to the third-ranking.

Table 5. Lecturer-1 expertise resulted from the baseline method AdjTFIDF and AdjTFIDF+SBERT for title-abstract data and AdjTFIDF+FastText for full-text data. Lecturer-1 is an expert in social media analysis.

Title-abstract Data/ AdjTFIDF	Title-abstract Data/ AdjTFIDF+SBERT	Fulltext Data/ AdjTFIDF	Fulltext Data/ AdjTFIDF+FastText
sentiment analysis	sentiment analysis	natural language	natural language processing
twitter	social media	indonesian language	language translation
customer satisfaction	marketing	ann	hoare logic
news recommendation	social recommendation	svm	mapreduce algorithms
machine learning	machine learning	twitter	blogs
indonesian language	language translation	buzzer detection	collision detection
news tags	social tagging	covid 19	heap (data structure)
svm	support vector machines	customer satisfaction	quality assurance
natural language	natural language processing	digital bank	digital cash
buzzer detection	speech recognition	election prediction	failure prediction

For title-abstract data, the AdjTFIDF+SBERT method with EN-ID data shows the best results with values of 0.58, 0.67, 0.59, 0.67, and 0.78 for metrics P@3, P@1, NDCG@3, NDCG@1 and MRR, respectively. For full-text data, the AdjTDIDF+SBERT method with ID data shows the best performance for four metrics, namely P@1, NDCG@3, NDCG@1, and MRR with values of 0.58, 0.43, 0.58 and 0.69, respectively. When comparing title-abstract and full-text data for two types of languages, ID and EN-ID, the best performance is shown by the AdjTFIDF+SBERT method with EN-ID title-abstract data. Unlike title-abstract data, in full-text data, the more data collected, the greater the possibility of variations from the publication keywords data, where these important words do not necessarily correspond to terms in expertise. In the title-abstract data, the important words obtained will be more specific and limited to increase the word importance level related to the publication keywords. Table 7 shows the expertise results sample of Lecturer-2 for ID and EN-ID title-abstract and full-text data.

Table 6. Evaluation Results for ID dan ENID Data. Significant differences related to AdjTFIDF/AdjTFID+SBERT/AdjTFIDF+FastText are highlighted using † for $p < 0.05$

Data	Method	P@3	P@1	NDCG@3	NDCG@1	MRR
Title-abstract ID	AdjTFIDF	0.44	0.42	0.44	0.42	0.61
	AdjTFIDF+SBERT	0.42	0.33	0.4	0.33	0.56
	AdjTFIDF+FastText	0.28	0.42	0.31	0.42	0.57
Title-abstract EN-ID	AdjTFIDF	0.39	0.42	0.38	0.42	0.57
	AdjTFIDF+SBERT	0.58†	0.67	0.59†	0.67	0.78
	AdjTFIDF+FastText	0.44	0.58	0.46	0.58	0.67
Full-text ID	AdjTFIDF	0.31	0.5	0.35	0.5	0.62
	AdjTFIDF+SBERT	0.39	0.58	0.43	0.58	0.69
	AdjTFIDF+FastText	0.31	0.5	0.34	0.5	0.57
Full-text EN-ID	AdjTFIDF	0.33	0.25	0.32	0.25	0.51
	AdjTFIDF+SBERT	**0.42**	0.42	0.42	0.42	0.61
	AdjTFIDF+FastText	0.39	0.33	0.38	0.33	0.43

Table 7. Lecturer-2 expertise resulted in ID and ENID language data for the AdjTFIDF and AdjTFIDF+SBERT. Lecturer-2 is an expert in image or object detection.

Lang	Title-abstract Data/ AdjTFIDF	Title-abstract Data/ AdjTFIDF+SBERT	Fulltext Data/ AdjTFIDF	Fulltext Data/ AdjTFIDF+SBERT

ID	Transfer learning	Transfer learning	Transfer learning	Transfer learning
	Fine tuning	Learning settings	Fine tuning	Learning settings
	Cross validation	Cross-validation	Deep learning	Machine learning
EN-ID	Automatic detection	Object detection	Faster rcnn	Deep belief networks
	Image detection	Image segmentation	Detection	Object detection
	Tuberculosis detection	Shape analysis	Fine tuning	Learning settings

Based on the annotation task, we filter the term in ACM classification terms for further evaluation. Several terms in the ACM classification are not considered in the expertise categories, such as 'annotation', 'b-trees', 'best practices for media', and 'buffering'. Terms that are considered not to be expert terms by annotators that appear in the top ten rankings will be removed so that expertise terms at the bottom will fill the rankings above them. In all scenarios, the minimum number of expert terms obtained after filtering is one, so evaluation calculations are only done on the first rank. For AdjTFIDF+SBERT and AdjTFIDF+FastText methods, the precision@1, NDCG@1, and MRR values will be the same, as shown in Table 8.

Table 8. Evaluation Results after the ACM Term Filtering Process. Significant differences related to AdjTFIDF-AdjTFIDF+SBERT and AdjTFIDF-AdjTFIDF+FastText are highlighted using † for $p < 0.05$.

Data	Metric Evaluation	AdjTFIDF+SBERT+filteredTerm	AdjTFIDF+FastText+filteredTerm
Title-abstract ID	P@1, NDCG@1, MRR	0.58	0.58
Title-abstract EN-ID	P@1, NDCG@1, MRR	0.75	0.67
Full-text ID	P@1, NDCG@1, MRR	0.67	0.5
Full-text EN-ID	P@1, NDCG@1, MRR	0.67†	0.42
Title-abstract EN	P@1, NDCG@1, MRR	0.49	0.22
Full-text EN	P@1, NDCG@1, MRR	0.44†	0.47†

For title-abstract ID data, the AdjTFIDF+SBERT+filteredTerm and the AdjTFIDF+FastText+filteredTerm method are superior with the same score of 0.58 compared to AdjTFIDF with a value of 0.42 for P@1, NDCG@1. The most superior method for the title-abstract ENID data is AdjTFIDF+SBERT+filteredTerm, with a value of 0.75 for the P@1 and NDCG@1, slightly better than AdjTFIDF+SBERT with a score of 0.67. For full-text ID data, the most superior method is AdjTFIDF+SBERT+filteredTerm, with a value of 0.67 for P@1 and NDCG@1, slightly improved by 0.09 compared to AdjTFIDF+SBERT. Likewise, for full-text ENID data, the most superior method is AdjTFIDF+SBERT+filteredTerm, with a score of 0.67 for P@1, NDCG@1, and MRR, higher by 0.25 for P@1 and NDCG@1 and higher by 0.06 for the MRR metric. For title-abstract EN data, the AdjTFIDF+SBERT+filteredTerm method is the highest ranking with a score of 0.49 for the P@1, NDCG@1 metric slightly higher by 0.16 compared to the AdjTFIDF+SBERT method. For full-text EN data, the AdjTFIDF+FastText method is superior with a score of 0.47 for the P@1, NDCG@1 metric, 0.14 higher than AdjTFIDF+FastText.

This research demonstrates that AdjTFIDF+SBERT+filteredTerm with title-abstract data shows the best result for ID and ENID data, scoring 0.75 for P@1 and NDCG@1 metrics. Likewise, for EN data, the best is the AdjTFIDF+SBERT+filteredTerm method, with a score of 0.49 for abstract data, slightly higher by 0.07 compared to full-text data with the basic AdjTFIDF method. The experiment shows that using title-abstract data will produce better expert terms, considering that the title-abstract contains limited and essential words with a narrower scope. Titile-abstract data contains important information or keywords that represent the contents of the publication document. The terms in the abstract are identical or similar to those in the keywords. Selecting expertise terms in ACM classes can improve the performance of obtaining expertise information. For full-text data, the author's keywords will be more spread out and less specific than the abstract. The keywords obtained will be more general. SBERT has better context recognition capabilities of its Siamese network architecture, giving better embedding in more dependent vector space.

4. Conclusions

The main objective of this study is to develop a search for the expertise of Fasilkom UI lecturers based on scientific publication data. Our goal is to evaluate the performance of the expertise search system using the lecturer's name as a query to obtain the lecturer's expertise. The basic method used is the adjusted TFIDF method (AdjTFIDF), in which the terms calculated in the TFIDF formula are a series of keywords created by the authors in their publication documents. Keywords listed in the publication title gain more weight. Keywords in scientific publications vary widely, are sometimes technical, and do not always reflect expertise terms, so we propose keyword mapping to ACM classification classes. We used various embedding techniques to map the adjusted TFIDF keywords to the ACM classes, including BERT, MBERT, XML-RoBERTA, FastText, and SBERT. The terms in the ACM class are then re-selected, only the terms considered as expert terms are taken, and the results are evaluated. Based on the type of data language, the evaluation is divided into 1) English data and 2) Indonesian and mixed Indonesian-English data because the number of lecturers who have Indonesian

publication documents is much smaller than the number of lecturers who have English publications. We also compared the performance of two types of data: title abstract data and full-text data. Three annotators were instructed to give a score of 0 if the generated term did not match the lecturer's expertise and a score of 1 if the generated term matched the lecturer's expertise.

For the evaluation based on data language, the experimental results show that the AdjTFIDF+SBERT+filteredTerm method is equally superior for the first type of data, namely EN, and for the second type of language: ID and ENID. Using ENID data is superior to ID for the second type of data language. The use of abstract data is also much superior to the use of full-text data. Title-abstract-EN data usage scored 0.49 for the P@1 and NDCG@1 metrics. Utilization of title-abstract-ENID data obtained a score of 0.75 for both P@1 and NDCG@1 metrics. The evaluation results also show that the mapping process to the selected ACM classes can improve accuracy because it can map specific technical terms to terms that expertise indicates.

Refining the embedding model to produce a dependent vector space that is more in line with the task of expert retrieval would be better for further work. The best keywords from the adjusted TFIDF and the results of keyword mapping to the ACM classes can be combined as better expertise terms. Various variations of other embedding models can also be considered for further research so that they can be compared and proven experimentally. Retrieval expertise could be tested across diverse academic institutions and calculated on a larger dataset corpus to obtain better generalizations or represent more real-world word distributions.

References

[1] K. Balog, Y. Fang, M. de Rijke, P. Serdyukov, and L. Si, *Expertise Retrieval*. Now Foundations and Trends, 2012. doi: 10.1561/1500000024.

[2] R. Gonçalves and C. F. Dorneles, "Automated Expertise Retrieval: A Taxonomy-Based Survey and Open Issues," *ACM Comput. Surv.*, vol. 52, no. 5, Sep. 2019, doi: 10.1145/3331000.

[3] H. H. Lathabai, A. Nandy, and V. K. Singh, "Institutional collaboration recommendation: An expertise-based framework using NLP and network analysis," *Expert Syst. Appl.*, vol. 209, p. 118317, 2022, doi: https://doi.org/10.1016/j.eswa.2022.118317.

[4] E. Broek, A. Sergeeva, and M. Vrije, "When the Machine Meets the Expert: An Ethnography of Developing AI for Hiring," *MIS Q.*, vol. 45, pp. 1557–1580, Sep. 2021, doi: 10.25300/MISQ/2021/16559.

[5] B. Ju, "Does domain knowledge matter: Mapping users' expertise to their information interactions," *J. Am. Soc. Inf. Sci. Technol.*, vol. 58, no. 13, pp. 2007–2020, Nov. 2007.

[6] W. Lyu and J. Liu, "Soft skills, hard skills: What matters most? Evidence from job postings," *Appl. Energy*, vol. 300, p. 117307, 2021, doi: https://doi.org/10.1016/j.apenergy.2021.117307.

[7] R. Fulbright, "The Expertise Level BT - Augmented Cognition. Human Cognition and Behavior," D. D. Schmorrow and C. M. Fidopiastis, Eds., Cham: Springer International Publishing, 2020, pp. 49–68.

[8] A. Salatino, S. Angioni, F. Osborne, D. Reforgiato Recupero, and E. Motta, "Diversity of Expertise is Key to Scientific Impact: a Large-Scale Analysis in the Field of Computer Science," in *27th International Conference on Science, Technology and Innovation Indicators (STI 2023)*, Jun. 2023. doi: 10.48550/arXiv.2306.15344.

[9] C. S. Campbell, P. P. Maglio, A. Cozzi, and B. Dom, "Expertise identification using email communications," in *Proceedings of the Twelfth International Conference on Information and Knowledge Management*, in CIKM '03. New York, NY, USA: Association for Computing Machinery, 2003, pp. 528–531. doi: 10.1145/956863.956965.

[10] A. Kongthon, C. Haruechaiyasak, S. Thaiprayoon, and K. Trakultaweekoon, "Automatically Constructing Areas of Expertise Based on R&D Publication Data," in *2017 Portland International Conference on Management of Engineering and Technology (PICMET)*, 2017, pp. 1–6. doi: 10.23919/PICMET.2017.8125418.

[11] Q. T. Tho, S. C. Hui, and A. C. M. Fong, "A citation-based document retrieval system for finding research expertise," *Inf. Process. Manag.*, vol. 43, no. 1, pp. 248–264, 2007, doi: https://doi.org/10.1016/j.ipm.2006.05.015.

[12] Y.-B. Kang, H. Du, A. R. M. Forkan, P. P. Jayaraman, A. Aryani, and T. Sellis, "ExpFinder: A hybrid model for expert finding from text-based expertise data," *Expert Syst. Appl.*, vol. 211, p. 118691, 2023, doi: https://doi.org/10.1016/j.eswa.2022.118691.

[13] P. Chaiwanarom and C. Lursinsap, "Collaborator recommendation in interdisciplinary computer science using degrees of collaborative forces, temporal evolution of research interest, and comparative seniority status," *Knowledge-Based Syst.*, vol. 75, pp. 161–172, 2015, doi: https://doi.org/10.1016/j.knosys.2014.11.029.

[14] X. Song, J. Yan, Y. Huang, H. Sun, and H. Zhang, "A Collaboration-Aware Approach to Profiling Developer Expertise with Cross-Community Data," in *2022 IEEE 22nd International Conference on Software Quality, Reliability and Security (QRS)*, 2022, pp. 344–355. doi: 10.1109/QRS57517.2022.00043.

[15] S. Surakka and L. Malmi, "Delphi Study of the Cognitive Skills of Experienced Software Developers," *Informatics Educ.*, vol. 4, no. 1, pp. 123–142, 2005, doi: 10.15388/infedu.2005.08.

[16] J. E. Montandon and M. T. Valente, "Mining the Technical Skills of Open Source Developers," *An. do Concurs. Teses e Diss. (CTD); 2022 An. do XXXV Concurs. Teses e Diss.*, pp. 1–10, 2022, doi: 10.5753/ctd.2022.222910.

[17] T. Dey, A. Karnauch, and A. Mockus, "Representation of Developer Expertise in Open Source Software," in *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)*, 2021, pp. 995–1007. doi: 10.1109/ICSE43902.2021.00094.

[18] C. Li, W. K. Cheung, Y. Ye, X. Zhang, D. Chu, and X. Li, "The Author-Topic-Community model for author interest profiling and community discovery," *Knowl. Inf. Syst.*, vol. 44, no. 2, pp. 359–383, 2015, doi: 10.1007/s10115-014-0764-9.

[19] V. Kumar and N. Pedanekar, "Mining Shapes of Expertise in Online Social Q&A Communities," in *Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion*, in CSCW '16 Companion. New York, NY, USA: Association for Computing Machinery, 2016, pp. 317–320. doi: 10.1145/2818052.2869096.

[20] A. Askari, S. Verberne, and G. Pasi, "Expert Finding in Legal Community Question Answering BT - Advances in Information Retrieval," M. Hagen, S. Verberne, C. Macdonald, C. Seifert, K. Balog, K. Nørvåg, and V. Setty, Eds., Cham: Springer International Publishing, 2022, pp. 22–30.

[21] N. Ghasemi, R. Fatourechi, and S. Momtazi, "User Embedding for Expert Finding in Community Question Answering," *ACM Trans. Knowl. Discov. Data*, vol. 15, no. 4, Mar. 2021, doi: 10.1145/3441302.

[22] R. Menaha, V. E. Jayanthi, N. Krishnaraj, and N. Praveen sundra kumar, "A Cluster-based Approach for Finding Domain wise Experts in Community Question Answering System," *J. Phys. Conf. Ser.*, vol. 1767, no. 1, p. 12035, 2021, doi: 10.1088/1742-6596/1767/1/012035.

[23] N. Coulter, "ACM'S computing classification system reflects changing times," *Commun. ACM*, vol. 40, no. 12, pp. 111–112, Dec. 1997, doi: 10.1145/265563.265579.

[24] B. Rous, "Major update to ACM's Computing Classification System," *Commun. ACM*, vol. 55, no. 11, p. 12, Nov. 2012, doi: 10.1145/2366316.2366320.

[25] P. Rodríguez and A. Spirling, "Word Embeddings: What Works, What Doesn't, and How to Tell the Difference for Applied Research," *J. Polit.*, vol. 84, May 2021, doi: 10.1086/715162.

[26] S. Selva Birunda and R. Kanniga Devi, "A Review on Word Embedding Techniques for Text Classification BT - Innovative Data Communication Technologies and Application," in *Lecture Notes on Data Engineering and Communications Technologies*, J. S. Raj, A. M. Iliyasu, R. Bestak, and Z. A. Baig, Eds., Singapore: Springer Singapore, 2021, pp. 267–281.

[27] M. Koroteev, *BERT: A Review of Applications in Natural Language Processing and Understanding*. 2021. doi: 10.48550/arXiv.2103.11943.

[28] J. Seo, S. Lee, L. Liu, and W. Choi, "TA-SBERT: Token Attention Sentence-BERT for Improving Sentence Representation," *IEEE Access*, vol. 10, pp. 39119–39128, 2022, doi: 10.1109/ACCESS.2022.3164769.