ROMANIAN JOURNAL OF INFORMATION SCIENCE AND TECHNOLOGY
Volume 29, Number 1, 2026, 3–14
doi: 10.59277/ROMJIST.2026.1.01

Reference Recommendation for Large Language Models-Generated Text Using Deep Textual Representations

Shariq BASHIR[1,*]

[1]College of Computer and Information Sciences, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh, Saudi Arabia

Email: sbmirza@imamu.edu.sa*

* Corresponding author

Abstract. The increasing use of large language models (LLMs) in conversational systems raises concerns about the credibility and verifiability of the information they generate. These models often produce fluent and convincing responses that may lack factual basis or supporting evidence. To address this challenge, a reference recommendation approach is proposed to retrieve relevant citations from LLM-generated text. The proposed approach treats the LLM output as a query and employs Sentence-BERT to create deep contextual embeddings of documents. Retrieval performance is further enhanced by integrating Siamese and Triplet neural network architectures to model semantic similarity and applying a submodular scoring function to ensure relevance and diversity in recommended references. Performance tests on a domain-specific dataset demonstrate that the proposed approach outperforms traditional retrieval approaches and recent baselines in standard evaluation metrics, including F1 @ k and mean recurrence rank (MRR). This work offers a scalable and effective solution for improving the reliability of AI-generated content through evidence-based support.

Key-words: Conversational information system; information retrieval; information system; large language models; reference recommendation; Sentence-BERT.

1. Introduction

Searching for information using Large Language Models (LLMs) is becoming increasingly popular. This popularity is due to advances in natural language processing (NLP) [1]. NLP helps LLMs understand human language better. People like to interact with LLM chatbots [2] because they can understand and generate human-like text [3]. They handle complex user queries well

and provide answers. Although LLM chatbots have advanced NLP abilities. They can provide valuable information and assistance. However, they can make mistakes [4–6]. This poses a risk to the authenticity and reliability of their information. Misunderstandings could mislead readers with different levels of background knowledge. Therefore, it is crucial for users to verify the important information obtained from LLMs through reliable sources before making decisions or taking action based on its output. By incorporating cross-referencing or fact-checking from trusted sources, users can reduce the risk of misinformation. This ensures accuracy and reliability in their decision-making processes. This will maximize their benefits while minimizing risks in various applications, including education, healthcare, and customer service.

In non-AI chat forums like Reddit and Quora, users frequently use hyperlinks within comments [7]. These links enrich discussions and validate information, similar to academic citations, by connecting to external sources. This practice helps to justify claims, allowing users to verify facts and explore different perspectives. LLM chatbots can benefit from this hyperlink approach. However, there are no clear standards for when and where the information generated by LLMs should include hyperlinks. Research on reference recommendation for LLM-generated information can address this issue. Providing reliable references allows users to trace the origin of the information, helping them to verify its accuracy and credibility independently. This fosters transparency and accountability, enabling users to critically evaluate the information based on the criteria and reputation of the cited sources.

The approach proposed in the paper addresses this task through information retrieval, referred to as reference recommendation for LLM-generated text. It treats either the current answer or the entire user-LLM chatbot thread as a dynamic search query, aiming to retrieve the most relevant reference from a collection. This fosters more informed and enriching discussions, empowering users to verify claims and deepen their understanding of topics. The proposed approach leverages BERT (Bidirectional Encoder Representations from Transformers) [8, 9], an advanced learning model critical for analyzing AI-generated text due to its state-of-the-art performance in NLP. Unlike keyword-based methods, reference recommendation relies on contextual understanding. BERT's bidirectional architecture analyzes words in full context—considering both preceding and subsequent words—to capture meaning at the sentence, paragraph, or document level [10]. Additionally, BERT generates deep contextualized word representations, where each word's meaning is dynamically influenced by its surrounding text. This ensures accurate relevance scoring and natural language comprehension. Together, these features enable the retrieval of references precisely tailored to the topic at hand.

Main Contributions: This paper investigates the reference recommendation for the text generated by LLMs. This task helps users verify the information provided by LLM chatbots using trustworthy sources. To accomplish this, a BERT-based model is proposed for reference recommendation, along with a discussion on the creation of a scoring function to identify relevant references. The scoring function can be developed either heuristically or through supervised learning with a dataset containing verified reference lists. The proposed approach frames reference recommendation as the task of maximizing a submodular scoring function [11]. It employs Sentence-BERT (SentBERT) for efficient and scalable document scoring [12], significantly reducing processing time while maintaining accuracy. The paper further investigates the use of Triplet and Siamese networks to generate sentence embeddings and evaluate text similarity. These networks utilize citation graphs to define the relationships between information and references.

The rest of this paper is structured as follows: Section 2 reviews related work on citation

and reference recommendation systems. Section 3 describes the proposed framework, including the submodular scoring function and deep network architectures (Sentence-BERT, Siamese, and Triplet). Section 4 details the experiments, datasets, and results. Section 5 concludes the paper.

2. Related Work

The proposed approach aligns with conversational retrieval tasks [13], where users interact with retrieval systems through dialogues. To the best of current knowledge, no prior study has addressed the task of using LLM-generated text as queries to identify relevant references. Similar to this research, hyperlinks in online discussions function similarly to informal citations, connecting the conversation to external content [7]. Much research has focused on citation recommendations, primarily within academic contexts [14, 15]. The effectiveness of citation recommendation depends on accurately understanding the context in which the citation is used [16]. Recent advancements in content-based citation recommendations have employed advanced machine learning techniques such as deep representations learned by GRUs [17, 18], encoder-decoder models with attention mechanisms [19], LSTMs [20], and Transformers [21]. Additionally, some research focuses on recommending hyperlinks on social media platforms like Twitter, utilizing user data from previous publications, personal citation networks, or collaborative filtering methods. Many of these approaches also incorporate explicit networks or interaction graphs to suggest relevant content. The proposed retrieval approach does not rely on user-specific representations or predefined graph structures. A key advantage of this approach is its ability to generalize, enabling the seamless addition of new references to the candidate pool without the need for prior user interactions. This flexibility helps to effectively mitigate the cold-start problem, making the system robust and adaptable. Moreover, while scientific literature is typically well-written and semantically rich, making citation recommendations easier, text generated by LLMs is often informal, long, and sometimes incoherent, posing additional challenges for retrieval. The approach is not limited to scientific citations; instead, it is designed to collect and retrieve general references to enrich the conversation. Similar to this research, there has been significant research on news recommendation systems, which primarily focus on understanding a user's needs to generate personalized recommendations [22, 23]. While the approach shares similarities with these systems, it extends beyond merely recommending news articles. The goal is to retrieve general references, emphasizing the context of online forum discussions [24]. One of the closest related works in news recommendation involves using forum discussions to enhance recommendations, where comments from discussion forums serve as additional features for recommending news articles. Unlike these approaches, the proposed approach recommends a variety of content beyond news. Additionally, cited references are employed to enable scalable evaluation without requiring manual relevance judgments, thereby streamlining the process and improving efficiency.

3. Reference Recommendation Task

The retrieval approach proposed in this paper addresses the task through information retrieval. The objective is to predict which references are most relevant to the specific conversation at hand. Similar to an internet search, the generated answer from the LLM is treated as a query,

which is then matched against a large collection of documents. The goal is to identify the most relevant document for that specific query, considering the context of the entire conversation. This research presents several fascinating challenges. Extracting the right "query" from the LLM answer isn't just about keywords; it requires understanding the broader flow of the conversation and the user's intent. Conceptualized as a retrieval problem, a retrieved answer from the LLM is defined as a query q. This search process traverses a collection of documents $D = d_1, d_2, ..., d_n$, with the objective of identifying a set of documents R^* within corpus D that relevant to the query q. A document is deemed relevant if it supports valuable evidence to the generated answer of the LLM. Consider an abstract of paper as the "query document" and a vast repository of academic articles as the "corpus". The goal is to identify a subset of relevant articles to recommend as potential references. To select the most relevant references, a scoring function is employed. This function, represented by a function $f(R)$, can be built in two ways: by incorporating expert knowledge through hand-crafted rules, or by leveraging the power of machine learning. In the machine learning approach, training is conducted on a dataset in which each document is paired with an ideal set of references, denoted as $\widehat{R}$. A loss function, $L(\widehat{R}, R)$, is then used to evaluate how closely a predicted reference set R aligns with the ideal set $\widehat{R}$. The goal is to maximize the scoring function $f(R)$ within a given budget constraint, thereby optimizing the prediction in terms of the following optimization problem:

$$R^* = \text{argmax}_R \quad f(R) \text{ subject to } |R| \leq M \tag{1}$$

where M represents the maximum number of top references returned to the user. In this paper, values of M equal to 1, 3, and 5 are used.

Traditional information retrieval relies on ranking algorithms that score and select top references individually; however, it fails to capture the interconnectedness among potential references. This paper reformulates reference recommendation as an optimization problem using a submodular scoring function [11], which naturally aligns with the inherent structure of the task. It is considered : D – a set that contains all candidate references, d – an individual reference within the candidate set D, $R1$ – a recommendation list containing potential references (a subset of D), and $R2$ – another recommendation list (a different subset of D).

Submodularity captures the concept of "law of diminishing returns": the additional benefit (marginal gain) of adding a new reference to a recommendation list progressively decreases as the list grows larger. In simpler terms, including a reference that complements the information already present in the list becomes progressively less valuable compared to adding a reference that brings entirely new perspectives or information. Formally, submodularity is expressed as:

$$R1 \subseteq R2 \rightarrow [f(R2 \cup d) - f(R2)] \leq [f(R1 \cup d) - f(R1)] \tag{2}$$

By leveraging submodularity, it is possible to design optimization algorithms that explicitly take into account the complementarity between references. This enables the generation of more diverse and informative relevant reference recommendation, going beyond simply selecting the top-ranked individual references and leading to a more well-rounded and comprehensive reference list. Optimizing a submodular function is NP-hard, making exact solutions infeasible for large datasets [11]. To tackle this, approximation methods like those based on the greedy algorithm are commonly used. The greedy algorithm builds a solution step by step by maximizing the marginal gain at each step. Though not always optimal, it is efficient and comes with theoretical guarantees that ensure its performance is close to the best possible solution.

In essence, starting with an empty set (denoted as A), the greedy algorithm iterates through each possible element (denoted as $d \in D$) and assesses the discrete derivative (marginal gain) associated with adding that element to the current list. At each iteration, the algorithm selects the element that maximizes this gain and adds it to the solution set. This process continues until a desired number of elements (denoted as M) is reached, resulting in an approximate solution for the submodular optimization problem:

$$A_i = A_{i-1} \cup \left\{ \text{argmax}_{d \in D \setminus A_{i-1}} \Delta \left(d \mid A_{i-1} \right) \right\} \tag{3}$$

In [11], the authors proposed a set of submodular functions designed to balance relevance, coverage, and diversity in the recommendation list. Among these, the function that demonstrated the best empirical performance is a monotone submodular function that utilizes meta-information about the authors:

$$f(A) = \sum_{k=1}^{K} \sqrt{\sum_{l \in A \cap C_k} R_{kl}} \tag{4}$$

In the above question, C_k, for $k = 1 \ldots K$, denotes the clusters from a partition of the corpus, obtained by clustering based on author information. The term $R_{kl} \geq 0$ represents the reward for selecting the recommended citation from the k-th cluster. Due to the sublinear growth of the square root function, this formulation encourages selecting citations from diverse clusters.

3.1. Reference scoring

Although feature-based techniques such as BM25 and TFIDF have proven to be robust for document scoring and selection, there has been a growing focus on utilizing deep neural networks for document modeling and scoring due to recent advancements in the field. One might be tempted to directly apply deep neural networks to learn an optimal scoring function for reference recommendation. This approach strives to precisely evaluate the similarity between any two documents within a specified training dataset. However, this strategy faces scalability challenges because modern scientific literature collections contain vast amounts of documents, leading to an excessively large number of potential document pairs for training. This significantly increases the training data size and computational demands. Therefore, this work prioritizes efficiency by employing the deep textual representation presented in the following subsection. This representation significantly reduces the required training data size, as further detailed in subsequent sections.

3.2. Textual representation

While powerful techniques like pre-trained BERT models offer promising avenues for pairwise document comparisons, their computational demands pose significant challenges in terms of slowness and scalability issue for practical applications with large datasets [25]. This work addresses these challenges by adopting alternative approaches that prioritize efficiency and scalability:

Sentence-BERT: Rather than comparing pairs of documents directly, the approach leverages Sentence-BERT [12] to embed each document separately using a pre-trained BERT model.

This allows for significantly faster comparisons using simple functions like cosine or Euclidean distances.

DistilBERT: To further enhance efficiency and reduce memory requirements, the approach leverages DistilBERT, a lighter and faster variant of the original BERT model [26]. This choice significantly decreases the retrieval time for the most similar document pair, making the approach more suitable for practical applications.

Through these strategies, the proposed approach demonstrates the ability to achieve substantial reductions in processing time, transforming what would have taken tens of hours with pairwise BERT comparisons into a process that completes in just approximately ten seconds. This paves the way for utilizing the power of pre-trained models like BERT for reference recommendation in large-scale settings without compromising on practicality.

3.3. Network architecture

This paper explores the use of Sentence-BERT (SentBERT) for reference recommendation. SentBERT, built upon Siamese and Triplet networks, enables efficient generation of sentence embeddings for various tasks, including assessing similarity between texts [27, 28]. These networks have proven successful in learning semantic similarity between sentences and paragraphs [27, 29]. The proposed approach uses pre-trained SentBERT models that have been trained on large datasets. These models are then fine-tuned for the specific task of reference recommendation. For fine-tuning, the model is provided with samples of query documents and possible reference documents. When applied, the model calculates a similarity score between the user's query and each potential reference in the corpus.

Siamese Network: This network, essentially a single network divided into "twin" branches, processes two input vectors (query and candidate document) and generates output representations suitable for comparison using simple metrics like cosine or Euclidean distances [30]. To fine-tune the network and ensure its accuracy, the gap between the predicted similarity values and the desired values is minimized. The approach achieved this by minimizing the discrepancy, using either the Euclidean distance or the mean squared error metric. An illustration depicting the Siamese network built upon the sentence BERT encoder is provided in the supplementary document of the paper [35].

Triplet Network: This variant of the Siamese network utilizes the same network structure three times to compare the scores of positive and negative training samples [31] (see Fig. 1 (b) in the supplementary document of the paper [35]). Given a query document q, a relevant reference (positive) d_p, and an irrelevant reference (negative) d_n, the network is trained to ensure the predicted similarity between the query document and the relevant reference $sim(q, d_p)$ is higher than the similarity between the query document and the irrelevant reference $sim(q, d_n)$:

$$\text{triplet loss } = \max\left[sim\left(q, d_p\right) - sim\left(q, d_n\right) + 1, 0\right] \tag{5}$$

Employing a basic hinge function, a loss is triggered when the similarity score $sim\left(q, d_p\right)$ fails to surpass that of document $sim\left(q, d_n\right)$ by a margin of at least one unit. In cases where this condition is not met, the loss is adjusted to zero. By employing pre-trained SentBERT models and utilizing efficient network architectures such as Siamese and Triplet networks, accurate and scalable reference recommendation is pursued while addressing the computational limitations often associated with these techniques.

3.4. Network training

In this research framework, a positive instance is defined as a pair denoted as (q, d_p), where document q serves as the query document and d_p is acknowledged as a relevant document. Conversely, a negative instance is represented as (q, d_n), indicating that document d_n is known to be irrelevant to the query q. Based on previous research, the document embedding model is trained using specific training sets [12]. These sets are designed for Siamese networks and Triplet networks, ensuring tailored parameter learning. Specifically, for the Siamese network, pairs of $(q, d_p/d_n)$ are employed, where q is the query document, and d_p or d_n serves as a positive or negative match, respectively. Conversely, for the Triplet network, triplets of (q, d_p, d_n) are utilized, where q is the query document, d_p is a positive match, and d_n is a negative match. This approach ensures that both networks are trained effectively to distinguish between relevant and irrelevant documents, thus enhancing their ability to accurately assess document similarity and relevance.

3.4.1. Citation graph utilization

The inherent structure of the citation graph is exploited, where documents are represented as nodes and directed edges connect documents that reference one another. This graph allows us to define a distance score for the path $path(d_i, d_j)$ between document pairs as the shortest path length connecting them within the graph. Documents directly citing each other have a distance level of 1, while documents connected by an intermediary reference have a distance level of 2 and so on.

3.4.2. Positive sample selection

Positive training samples play a crucial role in guiding the model toward identifying relevant references. Document pairs with a distance of less than 5 are selected as positive examples, as they are likely to represent relevant references due to their direct or indirect citation relationships with the query document. For the Siamese network, the target similarity is defined as $sim(d_i, d_j) = \alpha^{path(d_i,d_j)-1}$ for positive examples, where α is a positive constant determined through validation. This constant influences the desired similarity score for known matching documents, guiding the network to prioritize these relevant references. The Triplet network uses the cosine distance it calculates to set a clear gap between the distances of two key elements: a chosen positive example (highly relevant) and a negative one (not relevant). This ensures the network prioritizes learning to distinguish truly relevant references. This ensures that the model prioritizes learning representations that place positive examples closer to the query document compared to negative examples in the embedding space.

3.4.3. Negative sample selection

Negative samples provide valuable context for the network to learn the distinction between relevant and irrelevant references. The following four strategies are employed for negative sample selection.

- **Random:** This baseline approach randomly selects documents not considered positive examples for the given query document.

- **Closest Neighbors:** Drawing upon the embedding space learned from a pre-trained model with random negatives, documents closest to the query document q in this space are selected. These "closest neighbors" represent documents potentially similar to the query but not necessarily relevant references.
- **MostDissimilar Neighbors:** Conversely, documents farthest from the query document in the embedding space are also selected. These contrasting examples assist the network in learning to distinguish irrelevant documents with significantly different content or meaning from the query.
- **Mix:** It contains mixture of closest neighbors and most dissimilar neighbors.

To ensure balanced training, the number of negative examples (samples) is set equal to the number of positive examples for each query document. This prevents the model from being biased towards either positive or negative examples during training.

4. Performance Evaluation

For this research task, the queries are the answers generated from LLMs. These queries are then used to rank documents in a collection, allowing us to retrieve most relevant recommendations. To assess the effectiveness of a ranking model, both a collection of documents and a set of queries with their corresponding relevant recommendations are required, serving as relevance judgments. These relevance judgments enable us to evaluate how well a ranking model performs in ranking relevant documents.

Dataset Construction: Due to the absence of a suitable benchmark dataset for this retrieval task, a custom test collection was constructed. A focused crawler was employed to gather documents from the internet that are relevant to predefined topics. The process began by identifying 140 subtopics under the broader themes of deep learning and network security. For each subtopic, the focused crawler was trained using a supervised approach. The subtopic was issued as a query to Google Scholar, and the top 20 research papers were downloaded and treated as relevant examples. Research papers were chosen over general webpages due to their rich citation structures. To collect negative examples, an irrelevant query term (e.g., *the*) was used, and the resulting documents were gathered. These positive and negative examples were then used to train the crawler using the LibSVM algorithm [32]. The trained crawler was subsequently deployed to retrieve up to 1000 research papers per subtopic.

In a typical reference recommendation task, human assessors manually evaluate the relevance of documents for each query. However, such an approach is resource-intensive. To address this limitation, an alternative strategy was adopted by leveraging the citations embedded within the collected research papers. These citations, selected by the original authors, serve as implicit relevance judgments. Additionally, the surrounding context of each citation can be interpreted as a pseudo-answer generated by a large language model (LLM). The objective is to retrieve these cited documents among the top-ranked recommendations. This approach provides a scalable and cost-effective means to evaluate system performance. For the tests, 1000 citations were randomly selected from the collection to serve as pseudo-answers.

The performance of the proposed approach was compared with TFIDF, BM25 [33], Citeomatic [34], and SubRef [11]. Furthermore, to assess the effectiveness of different retrieval models for the reference recommendation task, two established evaluation metrics were employed:

$F1@k$ score and Mean Reciprocal Rank (MRR) [34]. These metrics offer complementary insights into model performance, allowing for a well-rounded evaluation.

4.1. Sentence-BERT implementation

Two retrieval models based on the proposed approach were utilized. The first model, termed "SentBERT", is detailed in Section 3.4. This approach encodes the entire document into a compact vector space, enabling swift comparisons and retrieval of documents pertinent to the query. Recommendations are derived by identifying the top K documents whose vector representations closely match the query vector. The second retrieval model integrates the submodular optimization technique (SubRef) to refine references based on similarity scores from SentBERT. This retrieval model is termed "SentBERT+SubRef".

For model training, the following hyperparameters were used for the Triplet and Siamese networks. For both networks, the learning rate was set to 2×10^{-5}, while the batch size was fixed at 16 due to memory constraints. For the Siamese network, α was tuned within the range 0.1 to 0.9 in increments of 0.1, using the validation set for selection. The optimal value identified for α was 0.4. The number of training epochs is adjusted based on the distance between documents in the citation graph. As this distance increases (e.g., from direct citations to second- and third-level citations), the number of training examples grows rapidly. To maintain training efficiency and balance, the number of epochs is reduced for larger distances. This strategy enables the model to learn from a wide range of examples without excessive training time or risk of overfitting. For the Siamese network, the epochs per training for d = 1, d = 2, d = 3 and d = 4 are used with 30, 20, 10, and 5 respectively. For the Triplet network, the epochs per training for d = 1, d = 2, d = 3 and d = 4 are used with 25, 15, 7, and 4 respectively due to large number samples per epochs as compared to Siamese network.

4.2. Results

Table 1 presents a performance comparison between baseline retrieval models and two versions of the proposed approach. The first version, named "SentBERT", utilizes the Sentence-BERT model to embed the entire text of each document into a vector space. In this approach, the highest-ranked k documents are chosen as recommendations for each query. An alternative version, called "SentBERT+SubRef", integrates SentBERT's similarity scores with the SubRef submodular inference technique to determine the recommendations. The performance data in Table 1 reveals that SentBERT alone offers competitive results. It outperforms the three baseline models in the F1@5 metric and achieves second-best results across all other metrics. This indicates that SentBERT is effective in retrieving relevant documents, demonstrating its capability to understand and match query contexts efficiently. The optimal configuration for SentBERT was determined through a validation set, as outlined in Tables 2 and 3 of the supplementary document of the paper [35]. The "SentBERT+SubRef" surpasses all metrics, showing significant improvements over the runner-up models like Citeomatic. Notably, it improves Mean Reciprocal Rank (MRR) by 16 percentage points and enhances F1 scores at levels (F1@1, F1@3, and F1@5) by 21%, 21%, and 30%, respectively. These enhancements highlight the effectiveness of combining submodular inference with SentBERT, providing more accurate and diverse recommendations.

The selection of training examples is important for the success of the proposed approach. Training with all possible negative samples is impractical due to their exponential growth with

dataset size, which would result in highly imbalanced training sets. Similarly, the selection of positive examples greatly influences model performance. The validation results for both the Triplet and Siamese networks are available in Table 2 and Table 3 of the supplementary document of the paper [35].

Table 1. Effectiveness of different retrieval approaches on the test samples

Approach	F@1	F@3	F@5	MRR
TFIDF	0.068	0.060	0.041	0.232
BM25	0.098	0.091	0.072	0.272
SubRef	0.159	0.154	0.119	0.360

Approach	F@1	F@3	F@5	MRR
Citeomatic (Without Deep Learning Rank)	0.156	0.041	0.122	0.336
CiteomaticDRank (With Deep Learning Rank)	0.187	0.175	0.124	0.406
SentenceBERT	0.170	0.168	0.137	0.377
SentenceBERT+SubRef	**0.226**	**0.212**	**0.161**	**0.471**

These networks were trained using positive examples at different distance levels (1 to 4). Various strategies were employed for selecting negative samples, including random, closest, most dissimilar, and mixed selections. The results indicate that the Siamese network consistently outperformed the others across all scenarios. The optimal distance level was determined to be 2, which includes citations from directly-cited documents. This was paired with the most-dissimilar negative examples selection strategy to ensure a robust and balanced training process. The success of this approach underscores the importance of carefully curated training examples and strategic negative sampling in enhancing the model's performance. To conclude, the proposed approach demonstrates significant improvements over baseline models by effectively embedding document text and leveraging submodular inference. The careful selection of training examples and optimal configurations ensures robust performance, making the models highly effective for reference recommendations in diverse applications.

5. Conclusions

A new task of finding relevant references for text generated by LLMs is presented, framed as an innovative retrieval problem. To address this, a reference recommendation approach was developed that utilizes a deep document representation. This representation is generated by encoding each document using Sentence-BERT (SentBERT), a transformer-based text embedding method.

SentBERT was fine-tuned using positive and negative examples derived from various strategies within the citation graph. Additionally, a submodular scoring function was employed to predict the relevant reference list. The tests demonstrated that the proposed approach significantly outperforms all compared approaches, including a leading neural baseline, across all evaluation metrics.

Acknowledgments. This work was supported and funded by the Deanship of Scientific Research at Imam Mohammad Ibn Saud Islamic University (IMSIU) (grant number IMSIU-DDRSP2603).

References

[1] Y. CHANG, X. WANG, J. WANG, Y. WU, L. YANG, K. ZHU, H. CHEN, X. YI, C. WANG, Y. WANG et al., *A survey on evaluation of large language models*, ACM Transactions on Intelligent Systems and Technology **15**(3), 2024, pp. 1–45.

[2] J. K. KIM, M. CHUA, M. RICKARD and A. LORENZO, *ChatGPT and Large Language Model (LLM) chatbots: The current state of acceptability and a proposal for guidelines on utilization in academic medicine*, Journal of Pediatric Urology **19**(5), 2023, pp. 598–604.

[3] D. KALLA, N. SMITH, F. SAMAAH and S. KURAKU, *Study and analysis of ChatGPT and its impact on different fields of study*, International Journal of Innovative Science and Research Technology **8**(3), 2023, pp. 827–833.

[4] I. AMARO, A. DELLA GRECA, R. FRANCESE, G. TORTORA and C. TUCCI, *AI unreliable answers: A case study on ChatGPT*, Proceedings of International Conference on Human-Computer Interaction, Copenhagen, Denmark, 2023, pp. 23–40.

[5] M. BHATTACHARYYA, V. M. MILLER, D. BHATTACHARYYA and L. E. MILLER, *High rates of fabricated and inaccurate references in ChatGPT-generated medical content*, Cureus **15**(5), 2023, paper e39238.

[6] S. ZHENG, J. HUANG and K. C.-C. CHANG, *Why does ChatGPT fall short in providing truthful answers?*, arXiv preprint arXiv:2304.10513, 2023.

[7] K. ROS, M. JIN, J. LEVINE and C. ZHAI, *Retrieving webpages using online discussions*, Proceedings of ACM SIGIR International Conference on Theory of Information Retrieval, Taipei, Taiwan, 2023, pp. 159–168.

[8] J. DEVLIN, M.-W. CHANG, K. LEE and K. TOUTANOVA, *BERT: Pre-training of deep bidirectional transformers for language understanding*, arXiv preprint arXiv:1810.04805, 2018.

[9] M. V. KOROTEEV, *BERT: A review of applications in natural language processing and understanding*, arXiv preprint arXiv:2103.11943, 2021.

[10] B. T. KIEU, I. J. UNANUE, S. B. PHAM, H. X. PHAN and M. PICCARDI, *Learning neural textual representations for citation recommendation*, Proceedings of 25th International Conference on Pattern Recognition, Milan, Italy, 2021, pp. 4145–4152.

[11] T.-B. KIEU, S. B. PHAM, X.-H. PHAN and M. PICCARDI, *A submodular approach for reference recommendation*, Proceedings of 16th International Conference of the Pacific Association for Computational Linguistics, Hanoi, Vietnam, 2019, pp. 3–14.

[12] N. REIMERS and I. GUREVYCH, *Sentence-BERT: Sentence embeddings using Siamese BERT-networks*, arXiv preprint arXiv:1908.10084, 2019.

[13] P. OWOICHO, J. DALTON, M. ALIANNEJADI, L. AZZOPARDI, J. R. TRIPPAS and S. VAKULENKO, *TREC CAsT 2022: Going beyond user ask and system retrieve with initiative and response generation*, Proceedings of The Text Retrieval Conference, Gaithersburg, MD, USA, 2022, pp. 1–6.

[14] Z. ALI, I. ULLAH, A. KHAN, A. U. JAN and K. MUHAMMAD, *An overview and evaluation of citation recommendation models*, Scientometrics **126**(5), 2021, pp. 4083–4119.

[15] T. QIU, C. YU, Y. ZHONG, L. AN and G. LI, *A scientific citation recommendation model integrating network and text representations*, Scientometrics **126**(11), 2021, pp. 9199–9221.

[16] Y. DING, G. ZHANG, T. CHAMBERS, M. SONG, X. WANG and C. ZHAI, *Content-based citation analysis: The next generation of citation analysis*, Journal of the Association for Information Science and Technology **65**(9), 2014, pp. 1820–1833.

[17] T. BANSAL, D. BELANGER and A. MCCALLUM, *Ask the GRU: Multi-task learning for deep text recommendations*, Proceedings of 10th ACM Conference on Recommender Systems, Boston, MA, USA, 2016, pp. 107–114.

[18] C. PORNPRASIT, X. LIU, P. KIATTIPADUNGKUL, N. KERTKEIDKACHORN, K.-S. KIM, T. NORASET, S.-U. HASSAN and S. TUAROB, *Enhancing citation recommendation using citation network embedding*, Scientometrics **127**(1), 2022, pp. 233–264.

[19] T. EBESU and Y. FANG, *Neural citation network for context-aware citation recommendation*, Proceedings of 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, 2017, pp. 1093–1096.

[20] L. YANG, Y. ZHENG, X. CAI, H. DAI, D. MU, L. GUO and T. DAI, *A LSTM based model for personalized context-aware citation recommendation*, IEEE Access **6**, 2018, pp. 59618–59627.

[21] C. JEONG, S. JANG, E. PARK and S. CHOI, *A context-aware citation recommendation model with BERT and graph convolutional networks*, Scientometrics **124**(3), 2020, pp. 1907–1922.

[22] M. KARIMI, D. JANNACH and M. JUGOVAC, *News recommender systems – survey and roads ahead*, Information Processing & Management **54**(6), 2018, pp. 1203–1227.

[23] M. LI and L. WANG, *A survey on personalized news recommendation technology*, IEEE Access **7**, 2019, pp. 145861–145879.

[24] Q. LI, J. WANG, Y. P. CHEN and Z. LIN, *User comments for news recommendation in forum-based social media*, Information Sciences **180**(24), 2010, pp. 4929–4939.

[25] J.-S. LEE and J. HSIANG, *PatentBERT: Patent classification with fine-tuning a pre-trained BERT model*, arXiv preprint arXiv:1906.02124, 2019.

[26] V. SANH, L. DEBUT, J. CHAUMOND and T. WOLF, *DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter*, arXiv preprint arXiv:1910.01108, 2019.

[27] P. NECULOIU, M. VERSTEEGH and M. ROTARU, *Learning text similarity with Siamese recurrent networks*, Proceedings of 1st Workshop on Representation Learning for NLP, Berlin, Germany, 2016, pp. 148–157.

[28] B. HU, Z. LU, H. LI and Q. CHEN, *Convolutional neural network architectures for matching natural language sentences*, Proceedings of 28th International Conference on Neural Information Processing Systems, Montreal, QC, Canada, 2014, pp. 2042–2050.

[29] J. MUELLER and A. THYAGARAJAN, *Siamese recurrent architectures for learning sentence similarity*, Proceedings of 13th AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, 2016, pp. 2786—2792.

[30] J. BROMLEY, I. GUYON, Y. LECUN, E. SÄCKINGER and R. SHAH, *Signature verification using a "Siamese" time delay neural network*, Advances in Neural Information Processing Systems **6**, 1993, pp. 737—744.

[31] E. HOFFER and N. AILON, *Deep metric learning using triplet network*, Proceedings of Similarity-Based Pattern Recognition Workshop, Copenhagen, Denmark, 2015, pp. 84–92.

[32] Y. CHEN, *A novel hybrid focused crawling algorithm to build domain-specific collections*, Dissertation Abstracts International **68**(3), 2007, pp. 1–85.

[33] J. LIN, X. MA, S.-C. LIN, J.-H. YANG, R. PRADEEP and R. NOGUEIRA, *Pyserini: A Python toolkit for reproducible information retrieval research with sparse and dense representations*, Proceedings of 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Montreal, Canada, QC, 2021, pp. 2356–2362.

[34] C. BHAGAVATULA, S. FELDMAN, R. POWER and W. AMMAR, *Content-based citation recommendation*, arXiv preprint arXiv:1802.08301, 2018.

[35] S. BASHIR, Supplementary material of the paper S. BASHIR, *Reference Recommendation for Large Language Models-Generated Text Using Deep Textual Representations, Romanian Journal of Information Science and Technology, 2026*. [Online]. https://drive.google.com/file/d/12G8NT0ZlZ5rOYOdynK99OnadTq8uoBjq/view?usp=sharing.