This study focuses on a collection of tweets related to government services (e-government), which are preprocessed and transformed into a domain-specific corpus for query expansion. Conventional IR models struggle with unstructured and noisy content containing informal language and abbreviations, which reduces retrieval accuracy. To overcome these issues, this study proposes a hybrid query expansion (QE) model named ROCBERT-QE, which combines corpus-based retrieval (CBR) with bidirectional encoder representations from transformers (BERT). The model applies dual expansion, using corpus-based co-occurrence frequencies to capture lexical relationships and BERT embeddings to preserve semantic context. A domain-specific corpus consisting of 5,017 preprocessed tweets related to Indonesia's National Health Insurance (BPJS) was constructed, encompassing 6,215 unique terms that represent linguistic variation and informality in public discourse. Experimental results demonstrate that ROCBERT-QE outperforms baseline retrieval methods such as TF-IDF, BM25, and standard BERT. For single-word queries, Recall reached 0.8574 and Precision 0.8807, while for sentence-level queries, Recall was 0.8932 and Precision 0.9175. The synergy of frequency-based and contextual expansion enables effective handling of lexical noise and semantic ambiguity. The results confirm the scientific potential of combining corpus-based and transformer-based approaches in IR tasks involving unstructured content. Practically, ROCBERT-QE can be applied for real-time analysis of citizen discourse in e-government contexts, such as service evaluation, policy feedback, and early detection of public issues. The framework is scalable and adaptable to other domains with informal or multilingual data characteristics

Keywords: information retrieval, query expansion, corpus-based retrieval, BERT, social media, e-government

UDC 004.912:004.85:316.774
DOI: 10.15587/1729-4061.2025.340258

ENHANCING RETRIEVAL PERFORMANCE IN SOCIAL MEDIA WITH CORPUS-BASED QUERY EXPANSION USING BIDIRECTIONAL ENCODER REPRESENTATIONS FROM TRANSFORMERS

Roberto Kaban
Corresponding Author
Doctoral Program of Computer Science, Doctoral Student*
E-mail: robertokaban@students.usu.ac.id
Poltak Sihombing
Doctor of Philosophy (PhD) in Computer Science, Professor*
Syahril Efendi
Doctor of Philosophy (PhD) in Mathematics, Professor*
Maya Silvi Lydia
Doctor of Mathematics, Associate Professor*
*Department of Computer Science
Universitas Sumatera Utara
Dr. T. Mansur str., 9, Padang Bulan,
North Sumatera, Indonesia, 20222

Received 24.07.2025
Received in revised form 06.10.2025
Accepted date 14.10.2025
Published date 31.10.2025

How to Cite: *Kaban, R., Sihombing, P., Efendi, S., Lydia, M. S. (2025). Enhancing retrieval performance in social media with corpus-based query expansion using bidirectional encoder representations from transformers. Eastern-European Journal of Enterprise Technologies, 5 (2 (137)), 70–83. https://doi.org/10.15587/1729-4061.2025.340258*

1. Introduction

The rapid growth of social media platforms has fundamentally reshaped the way people communicate, exchange information, and interact with institutions [1, 2]. Every day, users generate massive volumes of short and informal messages that are highly dynamic in nature [3]. Unlike traditional documents, these posts are filled with abbreviations, creative spelling, emoticons, and context-dependent expressions [4]. As a result, social media has become one of the most complex and challenging environments for information retrieval (IR) [5]. The traditional paradigm of keyword-based search, which was effective for structured or semi-structured text, is often insufficient for handling this highly unstructured and noisy data [6]. Consequently, one of the major scientific issues in modern IR is how to adapt retrieval methods to the unique characteristics of social media content.

A key factor underlying the complexity of IR in social media is the quality of user queries. Queries are often extremely brief, ambiguous, and underspecified [4, 6]. For example, in the domain of healthcare e-government, a simple query such as "BPJS service" may refer to very different aspects: user registration, membership validation, claim submission at hospitals, service quality, or subsidy coverage. Furthermore, queries frequently include spelling errors, informal abbreviations, or inconsistent terminology [3], such as "bpjs clime" instead of "BPJS claim". This variability makes it difficult for retrieval systems to capture the intended meaning, leading to low recall and precision. From a scientific perspective, this situation highlights a fundamental limitation of conventional IR approaches: they fail to bridge the semantic gap between short, noisy queries and the vast amount of unstructured social media data [5].

Within the IR community, query expansion (QE) has long been recognized as an important research direction to address this problem [7]. QE enriches user queries with synonyms, semantically related concepts, or contextually relevant terms [8, 9], thereby increasing the likelihood that retrieval systems will capture the true intent of users. Over the years, numerous methods for QE have been explored,

ranging from statistical co-occurrence and probabilistic models [10], to knowledge-based approaches [11] and, more recently, deep learning techniques [12, 13]. These methods have consistently demonstrated improved retrieval effectiveness in different application domains [14, 15]. However, the scientific issue remains that most existing approaches were developed and tested in relatively structured environments, not in the fast-evolving, noisy, and context-dependent landscape of social media [16, 17]. This gap underlines the continuing relevance of QE research in the present era.

The importance of advancing QE methods becomes even clearer when considering domain-specific contexts such as e-government. Social media today is not only a space for personal expression but also a key platform for civic engagement [18]. Citizens increasingly rely on platforms like Twitter (X) to express concerns, report service issues, and provide feedback on government programs [16, 19]. For governments, analyzing these interactions offers valuable insights into citizen sentiment and the effectiveness of public services [20]. In a country like Indonesia, with over 167 million active social media users [21], this form of engagement represents both an enormous opportunity and a major challenge. On one hand, social media data contains rich signals about the quality of governance and healthcare services such as BPJS. On the other hand, without robust IR techniques, much of this information remains inaccessible, diluted by noise and ambiguity [22].

These scientific and practical considerations show why the research topic of query expansion for social media IR remains highly relevant today [7, 12]. Far from being outdated, the issue is gaining importance as online communication becomes increasingly central to governance, healthcare, and other critical sectors. Moreover, the continuous evolution of social media platforms, linguistic trends, and interaction modes ensures that the challenges associated with retrieval will persist and even intensify [23].

Therefore, research on the development of query expansion methods tailored to the unique challenges of social media data, particularly in domain-specific applications such as e-government healthcare, is highly relevant and urgently needed.

2. Literature review and problem statement

Information retrieval (IR) has advanced considerably with the introduction of query expansion (QE) methods, which reformulate user queries by adding synonyms, related concepts, or semantically aligned terms. This process improves recall and precision by addressing the problem of underspecified or ambiguous user queries. Previous studies have demonstrated the effectiveness of various QE approaches.

Most QE research has focused on English or other high-resource languages, leaving a gap for Indonesian, where social media discourse is dominated by informal language, abbreviations, and code-mixing [4]. Research [7] reviewed semantic QE approaches, emphasizing that adaptation to domain-specific and informal data remains an open challenge. Research [9] introduced proactive QE for streaming data using external sources, highlighting that context-aware expansions improve retrieval in real-time applications.

Research [10] applied probabilistic QE through multinomial naive Bayes to select expansion terms automatically from initial retrieval results. While it reduced noise in expansions, it was less adaptable to informal language commonly found on social media. Research [11] utilized ConceptNet for short-text semantic expansion, outperforming several baseline methods, but its reliance on curated knowledge graphs limited scalability in dynamic domains such as e-government. Research [12] proposed knowledge-aware QE using large language models, improving retrieval in text-rich datasets but facing limitations in noisy, short-text social media contexts.

Research [13] demonstrated that deep contextual models such as BERT can effectively capture semantic relationships. However, prior studies rarely integrated BERT with corpus-based QE for domain-specific social media content. Research [14] explored generative QE using artificially generated texts, demonstrating enhanced retrieval performance when integrated with a domain-specific corpus. Yet, the high computational cost and reliance on large-scale pretrained models reduce practical feasibility.

Research [15] proposed a semantic matching algorithm enriching queries with relevant concepts and entities; although effective in structured domains, it struggled in noisy, informal environments. Despite these algorithmic advances, several challenges remain when applying QE to informal, domain-specific contexts. These linguistic characteristics introduce significant noise, reducing retrieval accuracy. Moreover, although domain-specific corpora play a critical role in tailoring query expansion to specialized contexts, few studies have explicitly integrated corpus-based retrieval (CBR) into their expansion process. This limitation is particularly evident in e-government domains, where nuanced terminology and context-specific expressions are essential for understanding citizen feedback and public service discussions [19, 21]. Additionally, the study by [24] employed short-text conceptualization in microblogs, transforming user inputs into meaningful concepts using Probase as an external knowledge base. The results showed improved retrieval effectiveness; however, this approach depended heavily on external resources, limiting applicability for languages outside English.

Furthermore, while deep contextual models such as bidirectional encoder representations from transformers (BERT) have demonstrated enhanced performance in capturing semantic relationships, their integration with corpus-based query expansion remains underexplored. Most existing approaches either rely solely on lexical frequency or apply embeddings without domain adaptation, limiting their ability to handle the dynamic and noisy nature of social media data. As a result, retrieval systems continue to struggle with low recall and precision in extracting relevant information from large-scale, unstructured, and domain-sensitive datasets.

These limitations persist for both objective and subjective reasons. Objectively, there is a scarcity of large-scale, domain-specific corpora for Indonesian social media and e-government, which hampers the effective use of corpus-based methods. Subjectively, prior research has largely focused on demonstrating improvements in retrieval accuracy under idealized conditions, rather than addressing real-world challenges such as linguistic variability, domain specificity, and noisy data environments.

Therefore, the main unresolved problem emerging from the reviewed literature is the absence of a corpus-based query expansion method that leverages BERT's contextual embeddings to improve retrieval performance in social media and e-government domains, particularly in the context of the Indonesian language. There is a lack of an adaptive, cor-

pus-driven, and semantically informed query expansion model capable of addressing the linguistic and contextual challenges of social media while supporting domain-specific applications such as e-government. This gap motivates the present research to propose a corpus-based query expansion model enhanced with bidirectional encoder representations from transformers, aiming to systematically enrich user queries and improve retrieval performance in social media contexts.

3. The aim and objectives of the study

The aim of this study is to develop a query expansion (QE) model based on corpus content-based retrieval (CBR) to enhance Information Retrieval (IR) performance on social media, particularly Twitter (X). The proposed model is designed to address key challenges in user queries related to e-government services, including ambiguity, informality, and the lack of structured representation.

To achieve this aim, the following objectives were accomplished:

– to propose the architecture of the recall optimized corpus-based BERT – query expansion (ROCBERT-QE) model;

– to collect and preprocess social media text data related to e-government services by applying cleaning, normalization, and indexing, and to construct a domain-specific corpus from the most frequent terms;

– to implement corpus-based query expansion combined with BERT embeddings to enrich original queries with contextually relevant terms and enhance retrieval performance.

4. Materials and methods

4. 1. Object and hypothesis of the study

This study focuses on the development of a query expansion (QE) model based on corpus-based retrieval (CBR) integrated with contextual embeddings from BERT to enhance the performance of information retrieval (IR) in social media data. The object of the research is a collection of tweets related to government services (e-government), which are preprocessed and transformed into a domain-specific corpus serving as the foundation for query expansion.

The main hypothesis is that a corpus-driven QE model enriched with contextual embeddings can improve precision and recall compared to traditional keyword-based retrieval. The study assumes that tweets about government services contain recurring terms and domain-specific expressions that can be systematically captured in a corpus, and that contextual embeddings are able to represent the semantic relationships between these terms more effectively. Simplifications adopted include focusing solely on text-based content from Twitter without incorporating multimodal features such as images or metadata, and limiting the evaluation to retrieval performance metrics such as precision, recall, F1-score, and mean average precision (MAP).

4. 2. Preprocessing and data augmentation

The raw Twitter dataset related to government services contains considerable noise due to informal expressions, abbreviations, and dynamic language typical of social media platforms. To ensure data quality and usability for retrieval tasks, a series of preprocessing steps were applied. Text cleaning involved removing URLs, mentions, hashtags, emojis, and other non-linguistic symbols that do not contribute to semantic meaning. Normalization was then performed to standardize spelling variations and convert all text to lowercase. Stopword removal eliminated common function words such as “dan” or “yang,” which add little value to retrieval. Tokenization segmented sentences into individual word tokens, enabling effective feature extraction. Finally, the preprocessed text was indexed to provide a structured format that supports efficient search and retrieval.

In this study, data augmentation refers not to the synthetic generation of new data, but to enhancing representational richness within the dataset. The first strategy involved corpus construction, where a frequency-based domain-specific corpus was built from the collected tweets. This corpus captures the most frequent and contextually relevant terms in public discussions of e-government services and forms the knowledge base for query expansion. The second strategy employed semantic enrichment through bidirectional encoder representations from transformers (BERT), which generated contextual embeddings and semantically related term variations, complementing frequency-based expansion with deeper linguistic and contextual associations.

Through the combination of preprocessing and data augmentation, the dataset was transformed from raw, unstructured text into a semantically enriched resource. This transformation provides a solid foundation for building and evaluating the proposed query expansion model for improving information retrieval performance in the e-government domain.

4. 3. Dataset

The dataset for this study consists of tweets collected from X (formerly Twitter) between January 2024 and January 2025. To ensure domain relevance, the keyword “Layanan BPJS Kesehatan” was selected, as it represents a critical public service frequently discussed by citizens. A total of 5,017 tweets were retrieved using the official API.The raw dataset included noise such as retweets, duplicated posts, advertisements, and off-topic content. These were removed during preprocessing. The cleaned data then underwent text normalization and tokenization, yielding unique and authentic expressions of public discourse on BPJS Kesehatan services.

From this refined dataset, a domain-specific corpus was built by analyzing term frequencies and co-occurrence patterns. Additionally, contextual embeddings were generated using BERT to capture semantic relations and lexical variations beyond simple frequency counts. This dataset serves two purposes:

1) constructing the domain-specific corpus for query expansion;

2) providing a benchmark for evaluating retrieval performance in the e-government domain.

4. 4. Feature extraction procedure

The feature extraction and content analysis process utilized a transformer-based model, BERT (bidirectional encoder representations from transformers), to evaluate and extract features from each document in the corpus. BERT can understand word context within a sentence bidirectionally, producing more accurate word representations in the corpus, thereby improving accuracy in content-based retrieval. BERT transforms each tweet into a feature representation vector that reflects semantic relationships within the text. BERT processes input in the form of tokenized text.

Before being fed into the model, tweets are tokenized into subwords using WordPiece tokenization

$$T(d_i)=\{t_1,t_2,\ldots,t_n\}. \quad (1)$$

$T(d_i)$ – the result of tokenization of documents d_i and t_1, t_2, ..., t_n – the result tokens BERT processing. BERT also uses special tokens namely [*CLS*] as the start token and [*SEP*] as the separator token between sentence, so that results BERT input becomes

$$X=\begin{bmatrix}[CLS],\\ t_1,t_2,\ldots,t_n,[SEP]\end{bmatrix}. \quad (2)$$

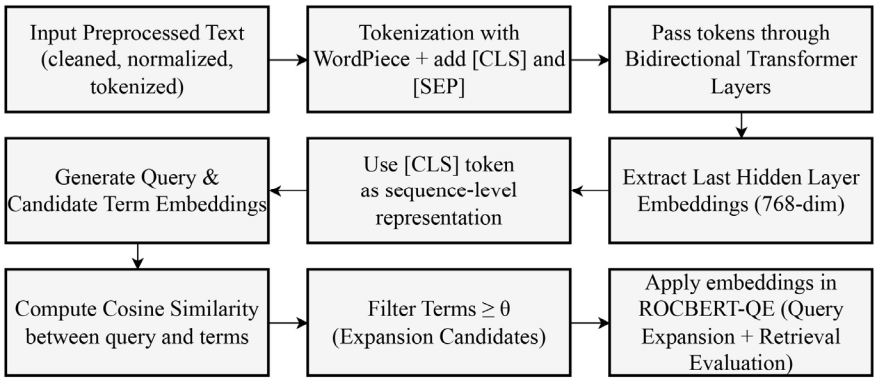

Fig. 1. Feature extraction procedure using IndoBERT

BERT uses Self-Attention Mechanism, which is calculated with

$$Attention(Q,K,V)=softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \quad (3)$$

where, *Q, K, V* – matrix representation of document tokens. d_k – dimensions from vector feature in hidden layer. *softmax* used for count weight attention between tokens. After the self-attention process, the representation feature for every document d_i obtained from hidden states final from the [*CLS*] token, which reflects overall context Documents

$$f_i=h_{[CLS]}^{(L)}, \quad (4)$$

where $h_{[CLS]}^{(L)}$ – hidden state final from tokens [*CLS*] on the *L* layer in the BERT model, f_i – features used for representation document. The final result is

$$F=BERT(C)=\{f_1,f_2,\ldots,f_k\}, \quad (5)$$

where, –a set of features generated from the embedding process by [BERT], f_i is vector features and documents d_i and k – the total number of features extracted.

In this study, IndoBERT (indobenchmark/indobert-base-p1) is used as the implementation of *BERT*. IndoBERT is pretrained on large Indonesian corpora, making it more suitable for processing queries and documents in Indonesian than multilingual *BERT*. Preprocessed text is tokenized using WordPiece, and special tokens such as [*CLS*] are added. The sequence is then passed through IndoBERT's bidirectional transformer layers. Embeddings are extracted from the last hidden layer, with two pooling strategies considered:

1) [*CLS*] token pooling;

2) mean pooling across token embeddings.

The [*CLS*] token is chosen for its compact sequence-level representation.

The resulting 768-dimensional embeddings serve as semantic representations of both queries and candidate terms. Cosine similarity is applied to measure semantic closeness, with terms exceeding a similarity threshold. Fig. 1 illustrates the feature extraction procedure using IndoBERT, from tokenization through embedding generation to its application in ROCBERT-QE.

The diagram shows the feature extraction procedure using IndoBERT. Input text is tokenized with the WordPiece algorithm, passed through IndoBERT, and embeddings are extracted from the last hidden layer using the [CLS] token. The resulting 768-dimensional vectors are then applied in ROCBERT-QE for query expansion and retrieval evaluation.

4. 5. Evaluation metrics

The evaluation of retrieval performance in this study employs several standard information retrieval (IR) metrics, namely precision, recall, F1-score and mean average precision (MAP). These metrics are widely used to provide a comprehensive assessment of retrieval effectiveness, both at the level of individual queries and across the entire system.

Precision (*P*) measures the proportion of retrieved documents that are relevant to the query

$$P=\frac{TP}{TP+FP}. \quad (6)$$

A high precision value indicates that most of the retrieved documents are indeed relevant, reflecting the system's ability to minimize false positives.

Recall (*R*) measures the proportion of relevant documents that are successfully retrieved from the total number of relevant documents available in the collection

$$R=\frac{TP}{TP+FN}. \quad (7)$$

A high recall value indicates that the system can retrieve the majority of relevant documents, minimizing false negatives.

*F*1-score represents the harmonic mean of precision and recall

$$F1=2\times\frac{P\times R}{P+R}. \quad (8)$$

This metric is particularly useful when a balance between precision and recall is required, ensuring that the system is not overly optimized for only one of them.

Mean average precision (*MAP*) provides a global measure of retrieval effectiveness across multiple queries. For a single query, the average precision (*AP*) is defined as

$$AP=\frac{1}{R}\sum_{k=1}^{n}P(k)\times rel(k), \quad (9)$$

where $P(k)$ – the precision at cut-off rank, $rel(k)$ – an indicator function (1 if the document at rank k is relevant, 0 otherwise), and R – the number of relevant documents.

For multiple queries, *MAP* is calculated as

$$MAP=\frac{1}{Q}\sum_{q=1}^{Q}AP(q). \quad (10)$$

MAP accounts for both relevance and rank position of documents, offering a more comprehensive evaluation than precision or recall alone.

4. 6. Experimental setup

The experimental setup in this study was designed to rigorously evaluate the performance of the proposed ROCBERT-QE model in comparison with baseline retrieval methods such as TF-IDF, BM25, and BERT without query expansion. All implementations were carried out using Python 3.10 as the main programming language. The Hugging Face Transformers library was employed for embedding extraction, while Scikit-learn supported the implementation of TF-IDF and BM25, as well as the computation of similarity measures and evaluation metrics. Additionally, a lightweight search engine – like prototype was developed to simulate the functionality of an information retrieval (IR) system, allowing queries to be submitted and results to be retrieved in real time.

The experiments were conducted on a personal computer running macOS Monterey, equipped with an Intel Core i5 processor operating at 1.6 GHz and 4 GB of DDR memory. Since the available hardware did not include GPU acceleration, all computations were executed on the CPU. This environment was adequate for the purposes of this research, as no large-scale training or fine-tuning of language models was required; the experiments focused exclusively on inference and retrieval processes.

The evaluation procedure consisted of three sequential steps. First, all candidate terms and documents in the corpus were preprocessed and indexed using TF-IDF, BM25, and BERT embeddings. Second, input queries were expanded with ROCBERT-QE by applying cosine similarity to the embeddings, thereby generating enriched query representations. Finally, the expanded queries were executed within the retrieval system, and the results were measured using the evaluation metrics outlined in Section 4.5, including precision, recall, F1-score and mean average precision (MAP). This setup ensured that the retrieval performance of ROCBERT-QE was assessed under consistent and reproducible conditions.

4. 7. Comparison with existing models

To further evaluate the effectiveness of the proposed ROCBERT-QE approach, its performance is compared against baseline retrieval models that do not employ query expansion, namely TF-IDF, BM25, and BERT. These models represent widely adopted retrieval techniques, ranging from classical frequency-based approaches (TF-IDF, BM25) to modern embedding-based methods (BERT).

The comparison is conducted on the same dataset of tweets related to "BPJS Kesehatan" services, ensuring a consistent evaluation context. All models are evaluated using the standard information retrieval metrics introduced in Section 4. 5, including precision, recall, F1-score and mean average precision (MAP). The evaluation covers four query scenarios: single-word queries (one word), two-word queries, three-word queries, and sentence-level queries (more than three words).

5. Results of ROCBERT-QE for social media retrieval

5. 1. Proposed recall optimized corpus-based BERT – query expansion (ROCBERT-QE) model architecture

The recall optimized corpus-based BERT – query expansion (ROCBERT-QE) model combines frequency-based corpus analysis with semantic embeddings to address ambiguous queries and noisy social media data. The architecture consists of two integrated components: a dynamic corpus construction module and a semantic expansion module. At the first layer, a Corpus module is constructed from a dataset of 5,017 tweets collected from social media platform X during January 2024 to January 2025, using the keyword "Layanan BPJS Kesehatan". After preprocessing steps such as tokenization and stopword removal, the frequency of each token is calculated. The Top-k most frequent terms are then selected to form a dynamic corpus, representing the vocabulary most commonly used by citizens in discussions of e-government services, especially "BPJS Kesehatan". This dynamic corpus ensures that the system reflects current and domain-specific language patterns, including informal terms, abbreviations, and slang commonly found in social media.

At the second layer, semantic embeddings based on BERT are applied to enrich the corpus-driven expansion. Both the initial user query and the candidate terms from the corpus are transformed into 768-dimensional vector embeddings. The connection between the query and the corpus is determined in order to add terms related to the query. Semantic similarity between queries and corpus terms is calculated using the cosine similarity method to identify terms that have a high level of relevance to the processed queries. The relationship between the query and corpus can be formally defined as a mapping function based on cosine similarity

$$R(Q,C)=\left\{q_i,f_j \mid \text{sim}(q_i,f_j)\geq\theta,\forall q_i\in Q,f_j\in F\right\}. \quad (11)$$

In this formula, $R(Q, C)$ represents the relation between the query Q and the corpus C, $\text{sim}(q_i, f_j)$ is the cosine similarity between the term q_i in the query and feature f_j in the corpus, and θ is the threshold of relevance. Terms which cosine similarity exceeds θ are considered relevant for expansion.

After obtaining the semantic connections between the query and the corpus, the most relevant terms are selected based on the highest similarity score

$$T=\left\{t_1,t_2,\ldots,t_p\right\}, \quad (12)$$

with

$$t_j=\arg\max_{f_j\in F}\text{Sim}\left(Q,f_j\right). \quad (13)$$

Here, T – the set of the most relevant terms extracted from the corpus for query expansion, and t_j – the term from the corpus that has the highest cosine similarity to the query.

Finally, the expanded query is obtained by adding these relevant terms to the original query

$$QE=Q'\cup T\ . \quad (14)$$

In this formula, Q' is the original query before expansion, T is the collection of terms added from the corpus, and QE is the expanded query ready for ranking and retrieval.

The integration of these two layers forms the ROCBERT-QE hybrid expansion mechanism. The final expanded query q' is obtained by combining the initial query q with the subset of corpus terms w_i that both appear frequently in the dataset and demonstrate high semantic similarity to the original query. Formally, this can be expressed as

$$q' = q \cup \left\{ \begin{array}{l} w_i \in Top-k\left(f\left(t\right)\right), \\ t \in \bigcup_{d \in D} Pre\left(d\right) \mid \frac{\phi\left(q\right) \cdot \phi\left(w_i\right)}{\phi\left(q\right) \cdot \phi\left(w_i\right)} \geq \\ \geq \theta \end{array} \right\}. \quad (15)$$

The formulation above defines the expanded query q' as the combination of the initial query q with candidate terms w_i selected from the corpus. The corpus is derived from social media data D, which is preprocessed using $Pre(d)$, including tokenization, stopword removal, and normalization, producing a set of tokens t. Each token's frequency is calculated using $f(t)$, and the top k most frequent terms ($Top-k(f(t))$) are chosen as candidates for query expansion. The semantic features of both the original query q and each candidate w_i are represented using BERT embeddings. Cosine similarity measures semantic relatedness, and only candidates with similarity scores above the threshold θ are incorporated into the expanded query. This approach ensures that the expanded query q' contains terms that are both statistically frequent in the corpus and semantically aligned with the user's original intent. By integrating frequency-based selection with BERT embeddings, the system effectively enhances recall while maintaining semantic relevance, enabling the retrieval of pertinent documents that may not explicitly match the initial query.

The overall workflow of the research is shown in Fig. 2 tweets are crawled, preprocessed, indexed, and expanded using both corpus-based and BERT-based approaches before retrieval through cosine similarity.

By dynamically constructing the corpus from social media data and enriching it with semantic embeddings, the ROCBERT-QE model provides an adaptive and context-sensitive query expansion method. This makes it particularly effective for handling ambiguous, short, and informal text found in tweets about BPJS Kesehatan services, where traditional retrieval methods often fail.

The detailed flow of ROCBERT-QE is illustrated in Fig. 3 it depicts sequential stages from corpus construction, semantic enrichment, query expansion, to retrieval.

The flowchart in Fig. 3 summarizes the end-to-end architecture of ROCBERT-QE, illustrating the sequential stages from raw tweet collection and preprocessing to corpus construction, semantic enrichment, and final query expansion for retrieval. This visual representation reinforces the textual and mathematical explanations above, and serves as the conceptual foundation for the implementation and experimental evaluation in the following sections.

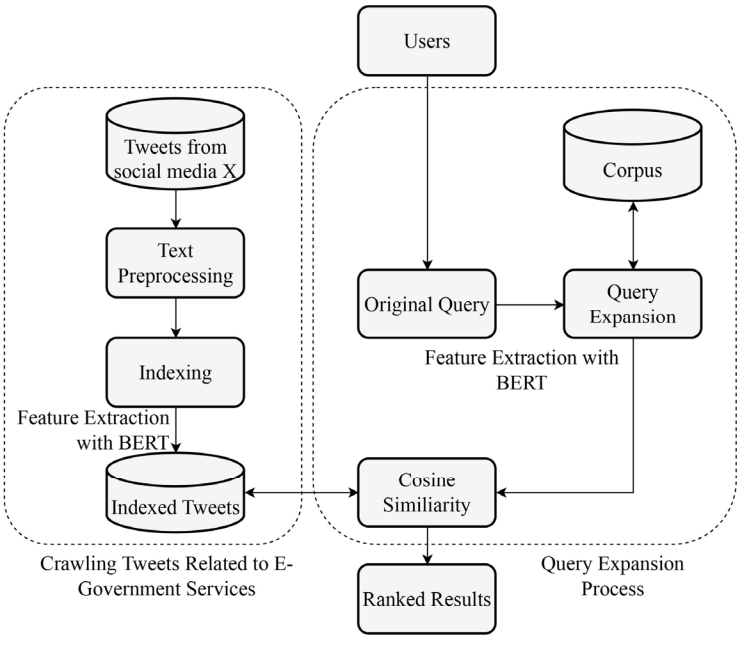

Fig. 2. Overall research architecture for query expansion and retrieval process

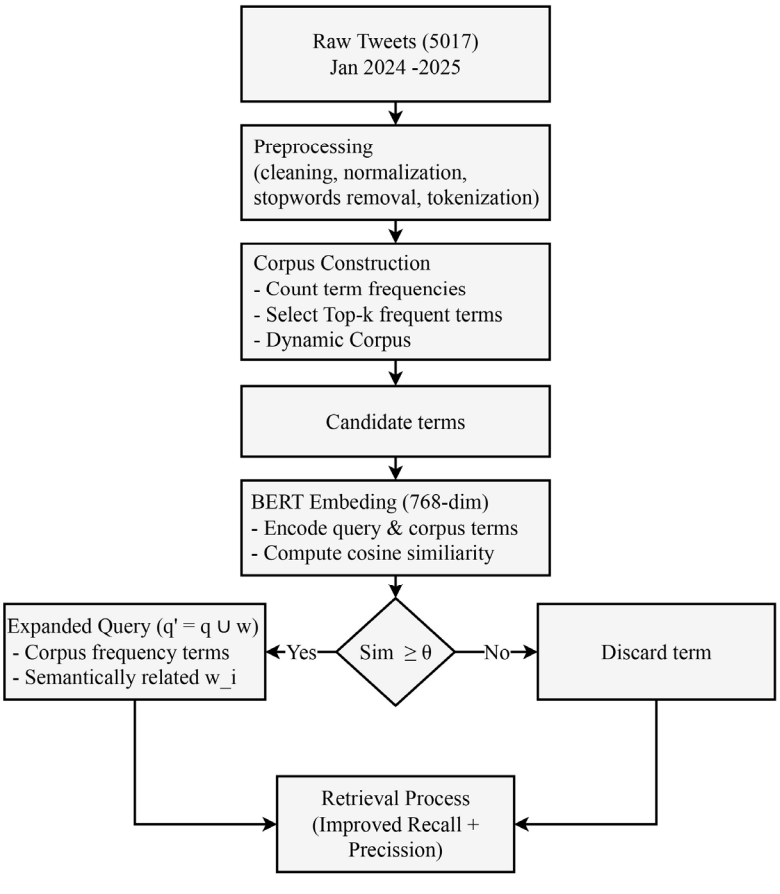

Fig. 3. Flowchart of ROCBERT-QE hybrid query expansion architecture

5. 2. Data collection and corpus construction

This stage focuses on cleaning, simplifying, and normalizing text data to facilitate more effective processing by the model. The raw data collected from social media platform X contains URLs, mentions, hashtags, symbols, numbers, and

emojis that may interfere with information retrieval and analysis. Data cleaning removes these elements to retain only meaningful textual content. For example, a tweet such as:

*"@dukcapil_id Tolong dong, KTP saya belum jadi juga padahal udah daftar sebulan lalu 😔🤲 #KTP #LayananPublik https://t.co/xyz*123*"*

after cleaning becomes:

"Tolong dong KTP saya belum jadi juga padahal udah daftar sebulan lalu".

In this process, mentions, hashtags, emojis, and URLs are eliminated, leaving only the relevant textual information. Following cleaning, tokenization is applied to divide the text into smaller units, or tokens, which facilitates accurate searching and analysis within the information retrieval system. For instance, the sentence:

"Saya sudah mengurus e-KTP sejak tiga bulan lalu, tetapi sampai sekarang belum jadi. Tolong dibantu!"

is tokenized into:

[*"Saya", "sudah", "mengurus", "e-KTP", "sejak", "tiga", "bulan", "lalu", "tetapi", "sampai", "sekarang", "belum", "jadi", "Tolong", "dibantu"*].

Stopword removal is then performed to eliminate frequently occurring words that carry little semantic meaning, such as *"yang", "dan", "atau", "di",* and *"ke"*. Removing stopwords enhances model efficiency by focusing on terms that contribute meaningfully to the content. After removing stopwords, the remaining tokens become:

[*"mengajukan", "permohonan", "SIM", "online", "diproses"*].

Stemming is applied to reduce words to their root forms by removing affixes, thereby consolidating variations of the same word into a single representation. For example, *"mengajukan"* is reduced to *"aju"* and *"diproses"* to *"proses"*. This ensures that words with the same semantic meaning are treated consistently.

Finally, normalization is performed to standardize the text, including converting all letters to lowercase and correcting informal or slang expressions to their formal equivalents, such as changing *"gak"* to *"tidak"*.

The result of these preprocessing steps is a clean, structured, and domain-specific corpus ready for feature extraction. A portion of the preprocessed data is presented in Table 1, showing that the text has been thoroughly cleaned, tokenized, stemmed, and normalized. This corpus forms the foundation for the subsequent stages of feature extraction and query expansion, ensuring that the system can effectively represent and process social media content related to e-government services.

After constructing the preprocessed corpus, feature extraction is carried out using IndoBERT. The text is first tokenized using the WordPiece tokenizer, which splits words into sub-word units to handle unknown words and variations in prefixes or suffixes. Special tokens such as [CLS], which marks the beginning of the text sequence, and [SEP], which separates segments, are added. Each token is then converted into a numerical ID and mapped to a dense vector representation provided by BERT.

The tokenized input is processed through multiple layers of the transformer encoder, where each token's representation is influenced not only by its own content but also by its relationship with other tokens in the sequence. The self-attention mechanism allows BERT to capture complex semantic dependencies and contextual nuances within the text.

The output of BERT consists of high-dimensional feature vectors for each token, which can be aggregated to represent the overall sentence. Common approaches include using the [CLS] token vector as a representation of the entire text, averaging vectors of all tokens, or combining outputs from several hidden layers to produce richer semantic representations. Each token is represented as a 768-dimensional vector that encodes its contextual meaning, capturing both positive and negative aspects of semantic information. These feature vectors form the numerical foundation for further stages, such as corpus-based query expansion and semantic enrichment of user queries.

Table 2 illustrates a sample of the extracted feature vectors for a subset of tweets, showing the transformation of textual content into structured numerical representations ready for retrieval and expansion tasks. The resulting vectors provide a comprehensive encoding of the textual corpus, enabling the system to perform semantic comparisons, measure similarity between queries and corpus terms, and support enhanced information retrieval performance.

Table 1

Tweets after preprocessing

No	conversation_id_str	created_at	text	...	username
1	1873712825673420000	Mon Dec 30 12:48:44 +0000 2024	temu buah dompet isi kartu tanda duduk surat izin kemudi stnk kartu bpjs sehat	...	seranim__
2	1769572162187530000	Mon Mar 18 03:50:55 +0000 2024	tinggal tetap sila ubah fktp sesuai alamat domisili baru tentu daftar minimal faskes terima kasih dewi	...	tanyarlfes
3	1786218098821070000	Fri May 03 02:15:55 +0000 2024	salam butuh layan sehat wilayah kotakabupaten fktp daftar sila faskes kunjung maksimal kali kunjung fktp dekat	...	tanyakanrl
...	...	...	...	...	...
5017	1760326826671640000	Wed Feb 21 15:33:15 +0000 2024	orang sakit gini kerja paksa tanggung asuransi sehat menteri keluarga pontang panting obat pake bpjs obat tanggung ikhlas allah	...	lapundery

Table 2

Feature extraction results using BERT

No	conversation_id_str	...	text	...	0	1	...	768
1	1873712825673420000	...	temu buah dompet isi kartu tanda duduk surat izin kemudi stnk kartu bpjs sehat	...	...	1.3745261	...	–0.01751
2	1769572162187530000	...	tinggal tetap sila ubah fktp sesuai alamat domisili baru tentu daftar minimal faskes terima kasih dewi	...	...	1.2654517	...	–1.07476
3	1786218098821070000	...	salam butuh layan sehat wilayah kotakabupaten fktp daftar sila faskes kunjung maksimal kali kunjung fktp dekat	...	...	0.039442778	...	–0.86812
...	...	...	...	...	...		...	
5017	1760326826671640000	...	orang sakit gini kerja paksa tanggung asuransi sehat menteri keluarga pontang panting obat pake bpjs obat tanggung ikhlas allah	...	...	0.46380496	...	–0.77263296

After completing the preprocessing stage, a domain-specific corpus is constructed from the cleaned social media texts. This corpus captures the most frequent terms used in discussions related to e-government services, particularly BPJS Kesehatan. Each term in the corpus is recorded along with its frequency and an example tweet in which it appears. This method ensures that the corpus reflects both the statistical importance of terms and their contextual usage in real social media conversations.

The resulting corpus consists of 6,215 unique terms. Each entry includes the term identifier, the term itself, its frequency in the dataset, and a sample tweet illustrating its use. Table 3 presents a subset of the corpus to demonstrate its structure and content.

Table 3

Resulting Corpus with Term Frequencies

id	term	frequency	example_tweet
1	sehat	5487	cover scalling masuk estetika kecuali parah ganggu...
2	bpjs	3641	kencing manis hipertensi kanker jantung stroke kom...
3	terima	2046	salam sehat sahabat mohon maaf sahabat pasti mengi...
...	...	...	...
6215	nanganin	1	nonaktif usaha kena tunggu bayar bayar kartu aktif...

This structured corpus serves as the foundation for the subsequent stages of feature extraction and query expansion. By linking each term to example tweets, the corpus maintains contextual relevance and semantic meaning, which is essential for improving information retrieval performance on social media texts regarding e-government services.

5. 3. Corpus-based query expansion with bidirectional encoder representations from transformers (BERT)

In this section, four types of queries were examined to evaluate the performance of the corpus-based query expansion using BERT. The queries differ in their length and structure, as listed below:

1. Single-word query.

The first evaluation involved a single-word query, “bpjs”. Using corpus-based expansion, the query included related terms such as “bpjs, bpjsnya, bpjstk”, allowing the system to retrieve additional contextually relevant documents. Four retrieval models – TF-IDF, BM25, BERT, and ROCBERT-QE – were compared. While TF-IDF and BM25 achieved perfect precision, TF-IDF had a low recall of 0.2431. BM25 improved recall to 0.717, resulting in a balanced F1-score of 0.8351. BERT provided a better trade-off with precision of 0.8695, recall of 0.8464, and F1-score of 0.8578. ROCBERT-QE outperformed all models, achieving precision of 0.8807, recall of 0.8574, F1-score of 0.8689, and MAP of 0.9212.

The comparative performance of all models is illustrated in Fig. 4.

The detailed performance metrics for each model are presented in Table 4.

Table 4

Precision, recall, F1-score, and MAP for each retrieval model evaluated with a single-word query

Method	Precision	Recall	F1-Score	MAP
TF-IDF	1.00000	0.24312	0.39115	0.95905
BM25	1.00000	0.71696	0.83515	0.95905
BERT	0.86950	0.84643	0.85781	0.86797
ROCBERT-QE	0.88075	0.85739	0.86891	0.92129

Table 4 shows that ROCBERT-QE effectively balances precision and recall, benefiting from corpus-driven query expansion to retrieve more relevant social media content.

2. Two-word query.

The second evaluation involved a two-word query: “layanan bpjs” (BPJS services). Corpus-based expansion extended the query to include terms such as “layanan bpjs, bpjsnya, centerlayanan, pelayananya, pelayananhal, bpjstk”. TF-IDF and BM25 again achieved perfect precision, but TF-IDF had low recall of 0.2431. BM25 improved recall to 0.7170, resulting in an F1-score of 0.8352. BERT achieved precision of 0.8725 and recall of 0.8494, yielding F1-score of 0.8608. ROCBERT-QE reached the highest performance with precision of 0.8895, recall of 0.8659, F1-score of 0.8775, and MAP of 0.9284.

The comparative performance of all models is illustrated in Fig. 5.

The detailed performance metrics for each model are presented in Table 5.

Table 5

Precision, recall, F1-score, and MAP for each retrieval model evaluated with a two-word query

Method	Precision	Recall	F1-Score	MAP
TF-IDF	1.0000	0.2431	0.3912	0.9591
BM25	1.0000	0.7170	0.8352	0.9591
BERT	0.8725	0.8494	0.8608	0.9094
ROCBERT-QE	0.8895	0.8659	0.8775	0.9284

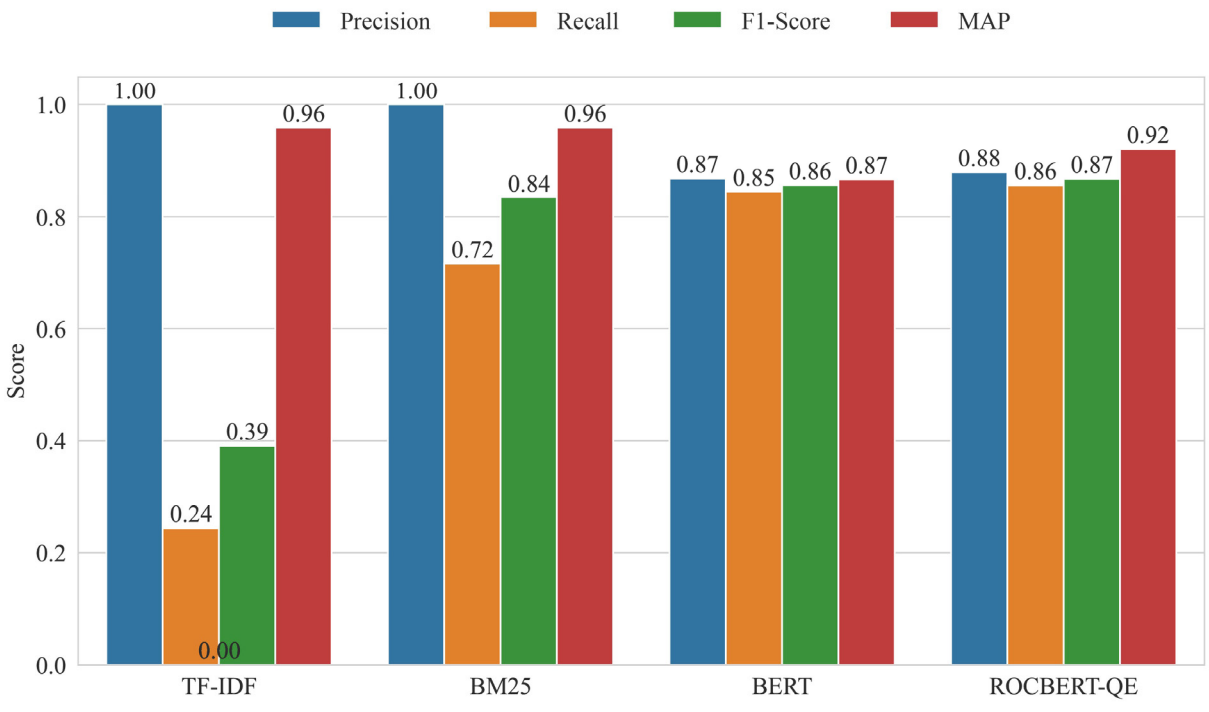

Fig. 4. Comparative performance of TF-IDF, BM25, BERT, and ROCBERT-QE using a single-word query ("bpjs")

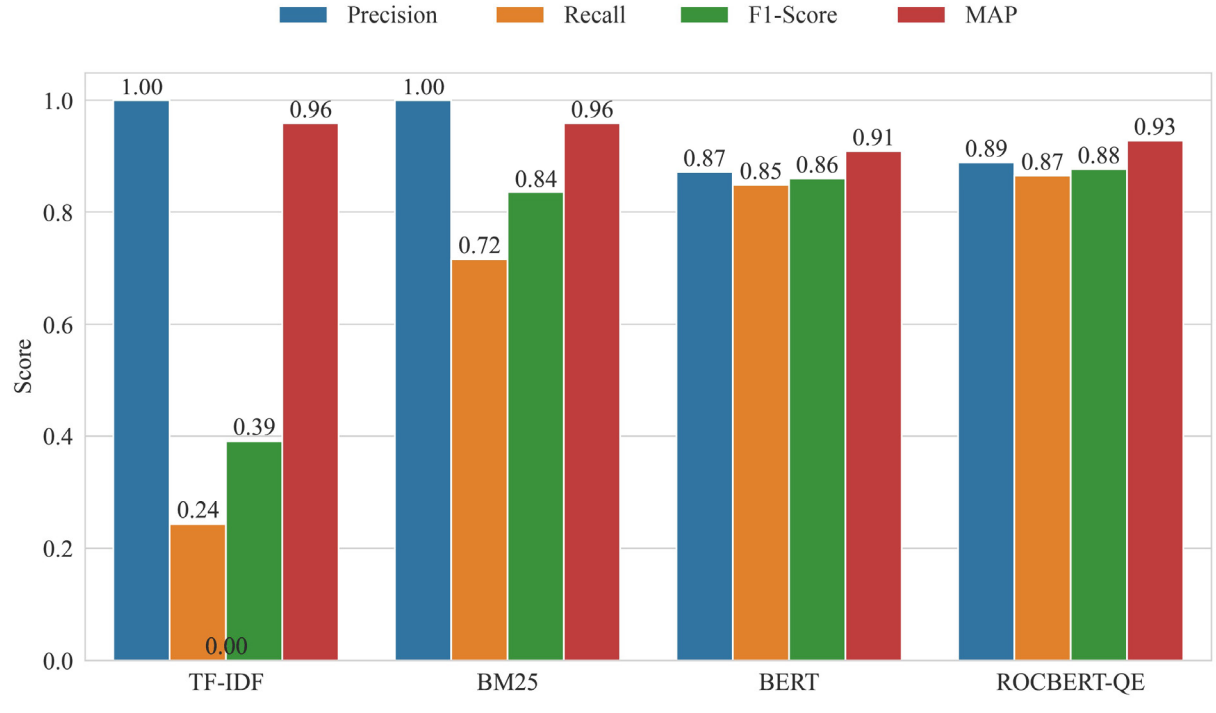

Fig. 5. Performance comparison of TF-IDF, BM25, BERT, and our approach for the two-word query "layanan bpjs"

ROCBERT-QE shows a clear advantage for short multi-word queries, maintaining high precision while improving recall.

3. Three-word query.

For a three-word query, "layanan bpjs kesehatan" (BPJS health services), corpus-based expansion included terms such as "layanan, bpjs Kesehatan, bpjsnya, kesehatanbpjs, ketenagakerjaankesehatan, centerlayanan, pelayananhal, kesehatanpolri, bpjstk pelayananya". TF-IDF and BM25 achieved perfect precision, but their recall remained low (0.24312 and 0.71696, respectively). BERT achieved precision of 0.87325, recall of 0.85009, and F1-score of 0.86151. ROCBERT-QE outperformed all, with precision of 0.89125, recall of 0.86761, F1-score of 0.87927, and MAP of 0.92519.

The comparative performance of all models is illustrated in Fig. 6.

The detailed performance metrics for each model are presented in Table 6.

The results show that ROCBERT-QE is effective for queries of increasing length, providing consistent improvement over standard models.

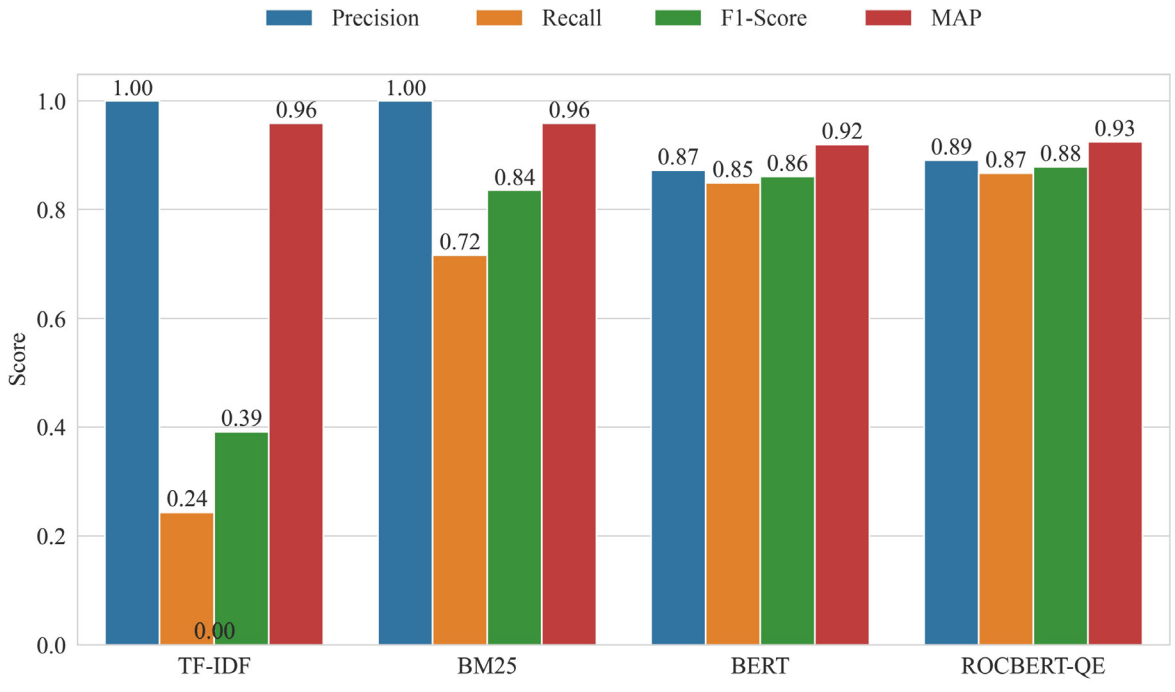

Fig. 6. Comparative performance of TF-IDF, BM25, BERT, and ROCBERT-QE using a three-word query ("layanan bpjs kesehatan")

Table 6

Precision, recall, F1-score, and MAP for each retrieval model evaluated with a three-word query

Method	Precision	Recall	F1-Score	MAP
TF-IDF	1.00000	0.24312	0.39115	0.95905
BM25	1.00000	0.71696	0.83515	0.95905
BERT	0.87325	0.85009	0.86151	0.92027
ROCBERT-QE	0.89125	0.86761	0.87927	0.92519

4. Long/sentence query.

The final evaluation involved a long query: "kenaikan tarif bpjs kesehatan berdampak terhadap rakyat miskin" (the increase of BPJS health fees affects poor citizens). Corpus-based expansion added multiple related terms. TF-IDF achieved high precision (0.96124) but very low recall (0.03018), resulting in a low F1-score of 0.05852. BM25 improved recall to 0.72426 with F1-score of 0.83819. BERT provided balanced performance with precision of 0.88275, recall of 0.85933, and F1-score of 0.87088. ROCBERT-QE achieved the best performance with precision of 0.91750, recall of 0.89316, F1-score of 0.90517, and MAP of 0.94604.

The comparative performance of all models is illustrated in Fig. 7.

The detailed performance metrics for each model are presented in Table 7.

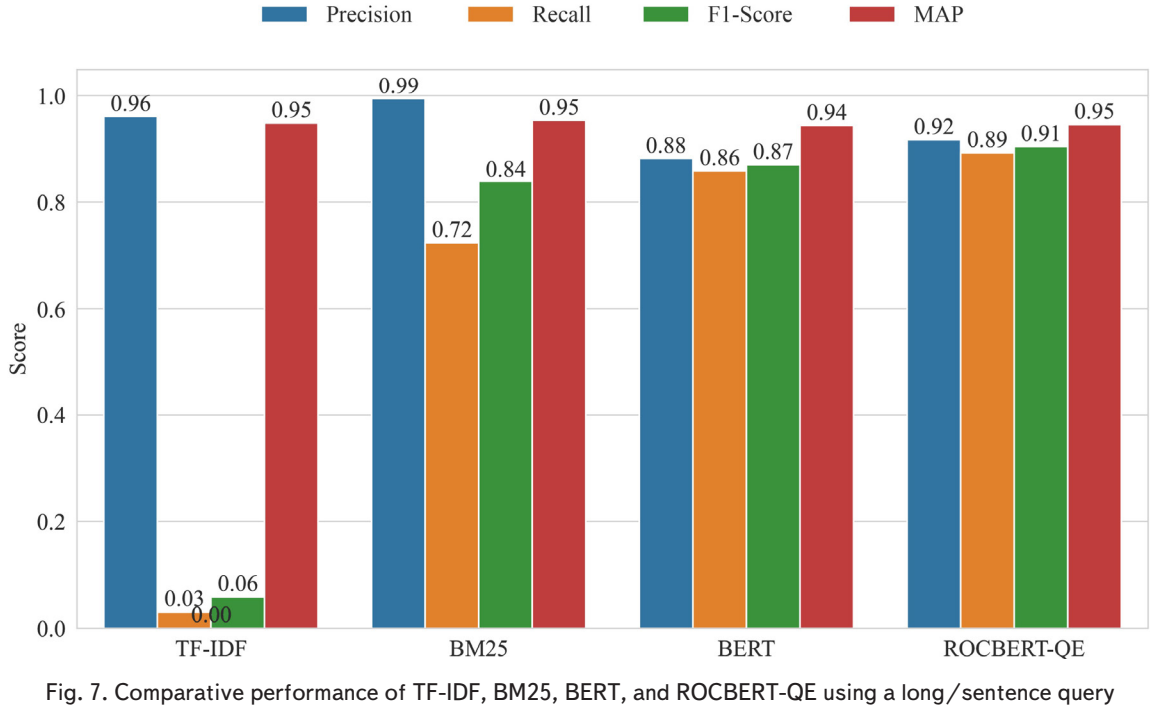

Fig. 7. Comparative performance of TF-IDF, BM25, BERT, and ROCBERT-QE using a long/sentence query

Table 7

Precision, recall, F1-score, and MAP for each retrieval model evaluated with a long/sentence query

Model	Precision	Recall	F1-Score	MAP
TF-IDF	0.96124	0.03018	0.05852	0.94865
BM25	0.99465	0.72426	0.83819	0.95409
BERT	0.88275	0.85933	0.87088	0.94434
ROCBERT-QE	0.91750	0.89316	0.90517	0.94604

The results indicate that ROCBERT-QE maintains a good balance between precision and recall, even for long and complex queries.

5. Recall and false negatives.

In addition to precision, recall is a crucial metric for evaluating retrieval performance, particularly in the context of short, informal, and noisy social media queries. Table 8 presents the recall and false negative rates for each model across different query types.

Table 8

Recall and false negative rates across models

Query	Model	Recall	False negatives
bpjs (1 word)	TF-IDF	0.2431	75.7% of tweets missed
	ROCBERT-QE	0.8574	14.3% of tweets missed
layanan bpjs (2 words)	BM25	0.7170	28.3% of tweets missed
	ROCBERT-QE	0.8659	13.4% of tweets missed
layanan bpjs kesehatan (3 words)	BERT	0.8501	14.9% of tweets missed
	ROCBERT-QE	0.8676	13.2% of tweets missed
long sentence	TF-IDF	0.0302	97% of tweets missed
	ROCBERT-QE	0.8932	10.7% of tweets missed

The table illustrates that corpus-based query expansion substantially reduces false negatives across all query types. Short queries, in particular, benefit from the inclusion of semantically related terms, while longer or multi-word queries achieve higher recall due to enriched contextual coverage.

6. Discussion of results for corpus-based query expansion with bidirectional encoder representations from transformers (BERT)

The presented ROCBERT-QE model is structured into two main layers that sequentially enhance query expansion. The first layer, the dynamic corpus construction layer, extracts the most frequent and domain-specific terms from preprocessed tweets related to "BPJS Kesehatan." This process captures informal expressions, abbreviations, and social media slang. The formal relationship between query terms and corpus terms is represented in formula (11), using cosine similarity to connect the user query with relevant corpus terms. The second layer, the semantic expansion layer, enriches queries by selecting corpus terms that are semantically aligned with the original query. The top-k relevant terms are identified according to formulas (12), (13) and incorporated into the query as described in formula (14). By integrating frequency-based selection and semantic similarity, formula (15) ensures that the expanded query contains terms that are both statistically significant and contextually relevant.

The overall workflow of ROCBERT-QE, including tweet crawling, preprocessing, indexing, and hybrid query expansion, is illustrated in Fig. 2, while Fig. 3 depicts the sequential flow from corpus construction to semantic enrichment, query expansion, and retrieval. This hybrid approach allows ROCBERT-QE to dynamically handle lexical noise and semantic ambiguity in social media texts, balancing statistical and contextual features to improve retrieval performance in domain-specific, informal, and noisy datasets.

As shown in Tables 1–3, the construction of a domain-specific corpus from preprocessed social media texts successfully captures informal expressions and terminology frequently used in "BPJS Kesehatan" related tweets. The cleaning, tokenization, stemming, and normalization steps illustrated in Table 1 ensure that only meaningful textual content is retained. The extracted feature vectors generated through BERT, as presented in Table 2, provide high-dimensional semantic representations that accurately encode contextual relationships among terms. As reflected in Table 3, this corpus provides a relevant foundation for query enrichment by highlighting dominant and contextually relevant terms in user-generated content.

As shown in Fig. 4–7 and Tables 4–7, the integration of corpus-based query expansion with BERT embeddings leads to measurable improvements in retrieval performance. For single-word queries, such as "bpjs," traditional frequency-based models like TF-IDF and BM25 maintain high precision but differ markedly in recall. Specifically, TF-IDF achieves perfect precision (1.0) but only 0.2431 recall, indicating limited coverage of relevant documents. BM25 improves recall to 0.717 while maintaining precision at 1.0. BERT embeddings provide a more balanced trade-off, improving recall to 0.8464 with precision of 0.8695. The ROCBERT-QE model outperforms all others, achieving recall of 0.8574 and precision of 0.8807, demonstrating that corpus-driven query expansion effectively enhances recall without compromising precision.

For two-word queries, such as "layanan bpjs" (BPJS services), TF-IDF and BM25 still show high precision (1.0) but recall remains limited (0.2431 and 0.717, respectively). BERT embeddings improve recall to 0.8494, yielding an F1-score of 0.8608. ROCBERT-QE further improves performance, achieving recall of 0.8659 and precision of 0.8895.

Three-word queries, for example "layanan bpjs kesehatan" (BPJS health services), provide additional context that allows semantic models to operate more effectively. Here, ROCBERT-QE reaches the highest F1-score of 0.8793 and MAP of 0.9252, while TF-IDF and BM25 continue to achieve perfect precision but with limited recall. For long sentence queries, such as "kenaikan tarif bpjs kesehatan berdampak terhadap rakyat miskin" (the increase of BPJS health fees affects poor citizens), ROCBERT-QE achieves precision of 0.9175, recall of 0.8932, and F1-score of 0.9052, significantly outperforming TF-IDF, which suffers from very low recall (0.0301) despite high precision (0.9612).

The performance of ROCBERT-QE can be contrasted with traditional and recent approaches in social media information retrieval. Classical models, such as TF-IDF and BM25, rely purely on keyword matching and frequency statistics. As observed in our results, these models maintain high precision but struggle with recall, especially for short or informal queries.

As shown in Table 8, ROCBERT-QE consistently reduces false negatives while maintaining competitive precision and recall across different query types. This consolidated comparison highlights that the integration of corpus-based query expansion with BERT embeddings provides a practical advantage in handling short and informal queries. Compared to baseline retrieval models, ROCBERT-QE reduces false negatives substantially, particularly for short queries, while sustaining or improving precision. For instance, in single-word queries such as "bpjs", the false negative rate decreased from 75.7% using TF-IDF to 14.3% with ROCBERT-QE, highlighting the practical advantage of integrating corpus-driven expansion with contextual embeddings.

Corpus-based query expansion introduces semantically related terms that may not appear in the original query. For instance, the single-word query "bpjs" is expanded into "bpjs, bpjsnya, bpjstk," allowing the system to retrieve relevant documents that do not explicitly contain "bpjs". This reduces false negatives, which is particularly important in social media and e-government information retrieval, where queries are often short and informal. The additional terms capture informal variations, abbreviations, typographical errors, and semantically related expressions common in social media discussions. For example, "bpjstk" refers to "BPJS Ketenagakerjaan", while "miskinekspresi" and "kemiskinansudah" reflect social contexts regarding the impact of BPJS policies on low-income populations. Learning from real social media corpora allows the retrieval system to handle non-standard language patterns that dictionary- or ontology-based expansions cannot capture.

Previous works have explored various query expansion strategies but with notable shortcomings. The study in [24] relied on Probase for short-text conceptualization, improving retrieval but restricted to English and external resources. A probabilistic QE approach using Multinomial Naive Bayes [10] reduced noisy expansions, yet lacked adaptability to informal language on social media. ConceptNet-based expansion [11] enhanced semantic matching but was limited in scalability for dynamic domains such as e-government. Semantic matching algorithms [15] showed effectiveness in structured domains, but failed to capture meaning in noisy and ambiguous contexts typical of Indonesian social media. Generative QE [14] achieved better performance with synthetic data, but at the expense of high computational cost. Unlike these methods, our proposed ROCBERT-QE combines corpus-based expansion with BERT contextual embeddings, enabling adaptive enrichment of queries with semantically aligned terms from domain-specific corpora. This allows the model to handle abbreviations, code-mixing, and informal discourse more effectively, reducing false negatives and consistently outperforming TF-IDF, BM25, and embedding-only baselines.

Overall, ROCBERT-QE consistently outperforms alternative methods across different query lengths, effectively balancing precision and recall. These findings confirm that combining corpus-driven query expansion with BERT embeddings provides a substantial advantage for social media information retrieval, particularly for informal, ambiguous, and unstructured queries.

Despite these advantages, several limitations must be acknowledged. First, the corpus is domain-specific and derived solely from Twitter, which restricts generalizability to other social media platforms or formal datasets. Second, the evaluation is limited to BPJS-related queries, meaning the results may not represent the broader e-government domain. Third, corpus-based expansion depends on frequency, which can introduce noise when high-frequency terms lack semantic relevance. Finally, the study focuses only on the Indonesian language, without testing multilingual or cross-lingual applicability.

Beyond these applicability limits, several disadvantages are inherent to the approach. One disadvantage is the reliance on computationally intensive BERT embeddings, making large-scale or real-time deployment less practical. Another disadvantage lies in the frequency-based query expansion mechanism, which occasionally introduces semantically irrelevant terms that reduce retrieval quality. These disadvantages can be mitigated by applying model compression techniques for efficiency and incorporating semantic filtering mechanisms to reduce noise in expanded queries.

Future research should address existing challenges, such as noise from irrelevant terms, dependency on corpus quality, and the high computational demands of BERT. Optimizing BERT through lighter models or model compression could improve efficiency. This approach could also be extended to multilingual contexts, supporting code-switching between Indonesian and English, and incorporating adaptive query expansion to follow evolving slang and trending terms on social media.

7. Conclusion

1. This study successfully proposed the architecture of the recall optimized corpus-based BERT – query expansion (ROCBERT-QE) model. The model integrates frequency-based corpus construction with BERT-based semantic embeddings to enhance query expansion for social media datasets. The architecture provides a structured foundation for improving retrieval accuracy, particularly in handling informal, ambiguous, and domain-specific language found in e-government-related tweets.

2. The collection and preprocessing of Twitter (X) data related to government services successfully produced a domain-specific corpus constructed from the most frequent and normalized terms. This corpus is distinct in its ability to capture informal expressions, abbreviations, and linguistic variations typical of social media, which are often overlooked in general-purpose corpora. By addressing this lexical gap, the corpus enhances the alignment between user queries and document representations. Compared to a generic corpus, it provided higher term coverage and improved vocabulary normalization accuracy, thereby offering a stronger foundation for effective query expansion in the government-related domain.

3. The implementation of ROCBERT-QE improved retrieval effectiveness compared to TF-IDF, BM25, and standard BERT. This proposed approach is distinguished by its ability to generate contextually relevant query expansions, reducing irrelevant matches and improving semantic precision. The enhancement was quantitatively demonstrated across different query types. For single-word queries, the model achieved an F1-Score of 0.8689 and MAP of 0.9212, outperforming BERT (F1-Score 0.8578, MAP 0.8679) and TF-IDF (F1-Score 0.3912). In the case of two-word queries, our method obtained an F1-Score of 0.8775 and MAP of 0.9284, surpassing BM25 (F1-Score 0.8352). For three-word queries, the system again led with an F1-Score of 0.8793 and

MAP of 0.9252. When handling longer queries in sentence form, the proposed framework reached an F1-Score of 0.9052 and MAP of 0.9460, exceeding BERT (F1-Score 0.8708, MAP 0.9443) and BM25 (F1-Score 0.8382, MAP 0.9541). The proposed model effectively addresses the limitations of recall and precision in traditional IR models while enhancing semantic matching through corpus-based query expansion.

Conflict of interest

The authors declare that they have no conflict of interest in relation to this study, whether financial, personal, authorship or otherwise, that could affect the study and its results presented in this paper.

Financing

The study was performed without financial support.

Data availability

Manuscript has data included as electronic supplementary material.

Use of artificial intelligence

The authors confirm that they did not use artificial intelligence technologies when creating the current work.

References

1. Ghanem, F. A., Padma, M. C., Alkhatib, R. (2023). Automatic Short Text Summarization Techniques in Social Media Platforms. Future Internet, 15 (9), 311. https://doi.org/10.3390/fi15090311
2. Sharma, S., Panda, S. P. (2023). Efficient information retrieval model: overcoming challenges in search engines-an overview. Indonesian Journal of Electrical Engineering and Computer Science, 32 (2), 925. https://doi.org/10.11591/ijeecs.v32.i2.pp925-932
3. Iamnitchi, A., Hall, L. O., Horawalavithana, S., Mubang, F., Ng, K. W., Skvoretz, J. (2023). Modeling information diffusion in social media: data-driven observations. Frontiers in Big Data, 6. https://doi.org/10.3389/fdata.2023.1135191
4. Benamara, F., Inkpen, D., Taboada, M. (2018). Introduction to the Special Issue on Language in Social Media: Exploiting Discourse and Other Contextual Information. Computational Linguistics, 44 (4), 663–681. https://doi.org/10.1162/coli_a_00333
5. Andreasen, T., Bordogna, G., Tré, G. D., Kacprzyk, J., Larsen, H. L., Zadrożny, S. (2024). The power and potentials of Flexible Query Answering Systems: A critical and comprehensive analysis. Data & Knowledge Engineering, 149, 102246. https://doi.org/10.1016/j.datak.2023.102246
6. Killingback, J., Zamani, H. (2025). Benchmarking Information Retrieval Models on Complex Retrieval Tasks. arXiv. https://doi.org/10.48550/ARXIV.2509.07253
7. Allahim, A., Cherif, A., Imine, A. (2025). Semantic approaches for query expansion: taxonomy, challenges, and future research directions. PeerJ Computer Science, 11, e2664. https://doi.org/10.7717/peerj-cs.2664
8. Massai, L. (2024). Evaluation of semantic relations impact in query expansion-based retrieval systems. Knowledge-Based Systems, 283, 111183. https://doi.org/10.1016/j.knosys.2023.111183
9. Alshanik, F., Apon, A., Du, Y., Herzog, A., Safro, I. (2022). Proactive Query Expansion for Streaming Data Using External Sources. 2022 IEEE International Conference on Big Data (Big Data), 701–708. https://doi.org/10.1109/bigdata55660.2022.10020577
10. Silva, S., Seara Vieira, A., Celard, P., Iglesias, E. L., Borrajo, L. (2021). A Query Expansion Method Using Multinomial Naive Bayes. Applied Sciences, 11 (21), 10284. https://doi.org/10.3390/app112110284
11. Chen, Z., Wang, J., Yang, X. (2022). A Concept Net-based semantic constraint method for query expansion. 2022 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), 906–913. https://doi.org/10.1109/wi-iat55865.2022.00147
12. Xia, Y., Wu, J., Kim, S., Yu, T., Rossi, R. A., Wang, H., McAuley, J. (2025). Knowledge-Aware Query Expansion with Large Language Models for Textual and Relational Retrieval. Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), 4275–4286. https://doi.org/10.18653/v1/2025.naacl-long.216
13. Fan, Y., Xie, X., Cai, Y., Chen, J., Ma, X., Li, X., Zhang, R., Guo, J. (2022). Pre-training Methods in Information Retrieval. IEEE Xplore. https://doi.org/10.1561/9781638280637
14. Claveau, V. (2020). Query expansion with artificially generated texts. arXiv. https://arxiv.org/abs/2012.08787
15. Kumar, R., Sharma, S. C. (2022). Hybrid optimization and ontology-based semantic model for efficient text-based information retrieval. The Journal of Supercomputing, 79 (2), 2251–2280. https://doi.org/10.1007/s11227-022-04708-9
16. Moreno-Ortiz, A., García-Gámez, M. (2023). Strategies for the Analysis of Large Social Media Corpora: Sampling and Keyword Extraction Methods. Corpus Pragmatics, 7 (3), 241–265. https://doi.org/10.1007/s41701-023-00143-0
17. Seo, W., An, H., Lee, S. (2025). A New Query Expansion Approach via Agent-Mediated Dialogic Inquiry. arXiv. https://doi.org/10.48550/arXiv.2502.08557

18. Shyrokykh, K., Girnyk, M., Dellmuth, L. (2023). Short text classification with machine learning in the social sciences: The case of climate change on Twitter. PLOS ONE, 18 (9), e0290762. https://doi.org/10.1371/journal.pone.0290762
19. Mutiarin, D., Wahyuni, H., Ismail, N. S. A., Kumorotomo, W. (2023). Social Media in Support of Indonesia's One Data Interoperability Process for Implementing Data Governance Policies. E3S Web of Conferences, 440, 03022. https://doi.org/10.1051/e3sconf/202344003022
20. Yulita, I. N., Al-Auza'i, A. F., Prabuwono, A. S., Sholahuddin, A., Ardiansyah, F., Sarathan, I., Djuyandi, Y. (2023). Bidirectional Long Short-Term Memory for Analysis of Public Opinion Sentiment on Government Policy During the COVID-19 Pandemic. International Journal of Advanced Computer Science and Applications, 14 (11). https://doi.org/10.14569/ijacsa.2023.0141189
21. Wahyudi, W., Loilatu, M. J., Omrilio, O. (2024). Capturing Political Conversation: An Analysis of Public Response on Social Media during Indonesian Election. Open Journal of Social Sciences, 12 (09), 163–182. https://doi.org/10.4236/jss.2024.129009
22. Abdelazim, H., Tharwat, M., Mohamed, A. (2023). Semantic Embeddings for Arabic Retrieval Augmented Generation (ARAG). International Journal of Advanced Computer Science and Applications, 14 (11). https://doi.org/10.14569/ijacsa.2023.01411135
23. Faggioli, G. (2023). Modelling and Explaining IR System Performance Towards Predictive Evaluation. ACM SIGIR Forum, 57 (1), 1–2. https://doi.org/10.1145/3636341.3636361
24. Wang, Y., Huang, H., Feng, C. (2017). Query Expansion Based on a Feedback Concept Model for Microblog Retrieval. Proceedings of the 26th International Conference on World Wide Web, 559–568. https://doi.org/10.1145/3038912.3052710