Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.078511

ARTICLE

# A Hybrid Approach for Query-Based Data Extraction Using Ensemble BERT Model with Walrus Optimization Algorithm

**Poluru Eswaraiah[1], Uddagiri Sirisha[2,*], Shaik Abdul Nabi[3], Revathi Durgam[4], Pallavi Malavath[5] and Gilakara Muni Nagamani[6]**

[1]Department of Computer Science and Engineering (Data Science), Vignan's Institute of Management and Technology for Women, Hyderabad, India
[2]Department of Computer Science and Engineering, Prasad V Potluri Siddhartha Institute of Technology, Kanuru, India
[3]Department of Computer Science and Engineering, AVN Institute of Engineering and Technology, Hyderabad, India
[4]Department of Computer Science and Engineering (Data Science), AVN Institute of Engineering and Technology, Hyderabad, India
[5]Department of Computer Science and Engineering (AI & ML), BVRIT Hyderabad College of Engineering for Women, Hyderabad, India
[6]Department of Computer Science & Information Technology, Koneru Lakshmaiah Education Foundation Deemed to be University, Green Fields, Vaddeswaram, India

*Corresponding Author: Uddagiri Sirisha. Email: sirisha.uddagiri@gmail.com

Received: 01 January 2026; Accepted: 20 April 2026; Published: 15 June 2026

**ABSTRACT:** The growing volume of digital text complicates the extraction of relevant information from unstructured data. Transformer models such as BERT, ALBERT, and RoBERTa are powerful, but they may face challenges in hyperparameter optimization and adaptation to new domains. To address this issue, a hybrid ensemble BERT model is suggested, optimized using the Walrus Optimization Algorithm (WaOA). The framework applies PCA to reduce dimensionality, ontology normalization, and K-means clustering to improve semantic comprehension. Experimental results on the SQuAD 2.0 and MS MARCO datasets show that the proposed model outperforms the baseline models. WaOA (Weighted Average of Attention) can improve convergence, reduce training time, and enhance prediction accuracy. The model also improves the semantic relevance of the extracted information. Attention maps visualize the model's focus on relevant query terms. The method enhances efficiency and cuts redundancy. It also provides a more generalized approach to different query types. The framework promotes consistent and reliable performance across different data conditions, including varying input formats and varying noise levels. It can be generalized to multilingual and domain-specific applications. Overall, the framework provides a scalable and reliable solution to real-world information extraction.

**KEYWORDS:** Query-based information extraction; ensemble BERT; walrus optimization algorithm; metaheuristic learning; PCA; K-means clustering; ROUGE; t-SNE; attention visualization

## 1 Introduction

The explosive spread of digital information across various fields has given people and entities previously unseen access to it. The problem of deriving meaningful, actionable knowledge from large volumes of unstructured text is a critical challenge, particularly given the abundance of such data. In the era of data-driven transformation, Information Extraction (IE) becomes increasingly significant, as it converts unstructured textual information into structured formats. Over the past 10 years, IE has emerged as a

significant field of study due to its applications in automation, knowledge discovery, and intelligent decision-making across numerous tasks, including information search, question answering, and text understanding. Recent work has addressed semantic link networks to extract meaningful word-level relations from text [1], contextual hypergraph-based models to improve extractive summarization with limited labeled data [2], and transfer learning approaches to adapt IE systems to new domains with little labeled data [3]. Besides, pretrained transformer encoders like BERT have proven to be highly effective for information retrieval tasks, enabling deeper contextual understanding and achieving state-of-the-art performance across various NLP benchmarks [4]. Simultaneously, question-answering systems that utilize deep learning architectures and models, including BERT, LSTM, and attention mechanisms, have also significantly increased the amount of data that can be extracted from unstructured text corpora [5].

Compared to pre-transformer architectures, the introduction of transformer-based models has fundamentally reshaped Natural Language Processing (NLP) and, by extension, modern Information Extraction (IE) systems [6,7]. BERT [8,9], ALBERT [10], DistilBERT [11], and ELECTRA [12] are examples of models that demonstrate impressive contextual knowledge through self-attention mechanisms that capture long-range, bidirectional relationships. Following the prosperity of BERT, researchers have designed variant transformers for domains of use, such as biomedical text mining, legal information extraction, and business document analysis [3]. Nevertheless, most transformer-based IE approaches still rely on fixed hyperparameter settings and single-model architectures. Furthermore, very few studies have systematically investigated the joint effects of feature dimensionality reduction, clustering, and adaptive optimization of transformer efficiency and contextual accuracy. These limitations reduce scalability and generalization, especially for low-resource or highly domain-specific corpora.

To address these shortcomings, this paper proposes a hybrid ensemble system that combines an ontology-guided semantic preprocessing step for concept normalization, K-Means clustering, Principal Component Analysis (PCA), and the Walrus Optimization Algorithm (WaOA) to improve query-based data extraction and summarization. This framework is unique because it has an adaptive optimization paradigm. In contrast to previous transformer hybrids based on optimization, such as PSO-BERT, GA-RoBERTa, or AOA-XLNet, which rely on single-stage or layerwise optimization, the proposed WaOA-enhanced Ensemble BERT uses a dual-phase experimentation-exploitation scheme inspired by walrus foraging patterns. This process continually updates the inertia and search direction of candidate hyperparameter solutions across multiple layers of the BERT ensemble, enabling the coordinated optimization of the learning rate, dropout, and attention depth. As a result, the model achieves a more stable convergence path, reduced overfitting, and better generalization across diverse query settings.

The proposed Ensemble BERT-WaOA architecture can deliver two fundamental results: increased semantic relevance and contextual accuracy of information extraction, and, at the same time, reduce the computational cost associated with feature compression and hyperparameter search. Even though transformer-based models have been widely used in text summarization and question answering, there are few comprehensive frameworks that combine semantic preprocessing with multi-layer, adaptive metaheuristic optimization. The study thus provides a unique methodological improvement by empirically assessing the optimized Ensemble BERT-WaOA system on the SQuAD 2.0 and MS MARCO benchmark datasets, demonstrating its capability as a high-fidelity, scalable system for query-driven information retrieval using domain-specific queries.

To achieve these goals, the paper analyzes the behavior of different transformer configurations in adaptive optimization and examines the effects of preprocessing steps, such as clustering and dimensionality reduction, on extraction accuracy and efficiency. In addition, we investigate whether a lightweight recurrent refinement on top of BERT can achieve stability across spans (i.e., make more consistent predictions for the

start and end of the answer to queries with varied query contexts, without the need to replace the transformer encoder). T These analyses collectively support a coherent and computationally efficient framework to improve the transformer-based information extraction in the real world.

## 2 Literature Survey

Information Extraction (IE) has long been a fundamental task in Natural Language Processing (NLP) and has been extensively studied under both rule-based and learning-based paradigms. Early methods mainly relied on manually constructed dictionaries and linguistic rules; this is the case with Tho et al. [13], who introduced the first dictionary-based IE systems, which tokenize documents and extract predefined entries when there is a lexicon match. Such techniques achieved reasonable accuracy, but building and maintaining large domain-specific dictionaries was time-consuming and labor-intensive.

Later developments introduced machine learning approaches that enable automatic feature learning and transfer to other domains. In the research works by Li et al. [1], Onan and Alhumyani [2], and Seow et al. [14], supervised and sequence learning models (Long Short-Term Memory (LSTM) networks) were used to model syntactic and semantic dependencies within the text. These models have greatly improved entity classification accuracy and reduced reliance on manual feature engineering. Machine learning-based IE techniques proved more adaptable to new domains and more efficient than dictionary-based methods, driving a paradigm shift in large-scale text processing.

In the field of IE, Named Entity Recognition (NER) has become a critical subtask that entails recognizing entities such as persons, organizations, and locations. The CoNLL-2003 shared task defined four major entity types: person, location, organization, and miscellaneous [11]. In practice, however, business and technical documentation worldwide often requires a more granular entity classification scheme to accommodate specialized terms. Li et al. [1] and Nguyen et al. [3] proposed solutions to the problem by suggesting fine-grained NER frameworks better able to handle domain-specific entities. Such methods have now been incorporated into larger NLP tasks such as question answering, information retrieval, and ontology building [8,10,15].

Transformer-based architectures have dramatically impacted the research on IEs. The first Transformer architecture, proposed by Vaswani et al. [16], demonstrated a self-attention-based system that replaced traditional recurrent and convolutional networks with a long-range structure and supports parallel computation. This invention significantly accelerated training procedures and performance in sequence transduction. With the introduction of Bidirectional Encoder Representations from Transformers (BERT) by Devlin et al. [9], pre-trained contextual encodings were introduced, which could be fine-tuned for downstream tasks, outperforming existing sequence models and demonstrating better generalization.

Following BERT's successful participation, various lightweight and focused versions have been proposed, including ALBERT [10], DistilBERT [11], and ELECTRA [12], each aiming to minimize model size or improve training speed. The models demonstrated state-of-the-art performance across a broad range of NLP benchmarks and serve as the basis for transformer-based IE systems. Devlin et al. [9] underlined that the bidirectional encoding and fine-tuning strategy used by BERT yields better results in contextual understanding tasks than unidirectional models like GPT, which provides a more comprehensive context for information extraction.

Additional work was on the use of transformers in domain-specific IE and question answering. Nguyen et al. [3] further refined BERT to extract structured information from business documents and demonstrated its robustness to domain adaptation. Tafjord and Clark [17] proposed MACAW, a multi-angle question-answering theory based on the UnifiedQA and T5 frameworks, which relied on multi-task transfer learning

to enhance reasoning and comprehension. Ghojogh and Ghodsi [18] conducted a more complex comparative study of attention mechanisms across various architectures, including BERT and GPT-3, and revealed structural and functional differences between bidirectional and autoregressive transformers.

NLP was extended to generative work with the introduction of autoregressive models like GPT-3 [19]. Floridi and Chiriatti [19] observed the power and limitations of GPT-3, noting its capacity to produce coherent text and its limited depth in semantic reasoning, and concluded that models with generation and semantic reasoning capabilities are necessary. Transformer encoder-decoder systems, including BART and T5, conversely, have shown even better results on extractive and abstractive summarization tasks, showing the power of transformer pre-training paradigms.

In addition to the fundamental transformer architectures, several supporting studies have extended the proposed information extraction and text interpretation. Alomari et al. [20] examined the concept of warm-starting to increase the diversity and novelty of abstractive summarization outputs, depending on the method used to initialize the summarization process, and found that the procedure is highly dependent on the method used to start the program. Parikh et al. [21] introduced automated utterance-generation mechanisms based on deep learning, which add to the overall knowledge framework by enabling more intricate neural models to produce relevant, coherent readings that conform to surrounding circumstances and context. In the field of semantic understanding, Seo et al. [22] studied Semantic role labeling across layers of the KR-BERT language model, emphasizing the importance of leveraging fine-grained syntactic and semantic structures within transformer layers for downstream extraction tasks. The algorithmic basis of the optimization strategy introduced in the current study is the Walrus Optimization Algorithm (WaOA) by Trojovsky and Dehghani [23], a bio-inspired metaheuristic for optimization problems, which serves as the basis for the algorithm implemented to address the optimization problems. Also, Shi and Lin [24] demonstrated that a plain BERT-based model applied to information extraction processes could attain levels of competitiveness with the best results in relation extraction and semantic role labeling tasks, which proved the multitask character of a BERT instantiation. Adnan and Akbar [25] have presented an ambitious analytical work on information extraction from unstructured, multidimensional big data, highlighting the need for scalable, flexible extraction patterns applicable to a variety of data modalities.

Other recent studies have examined type 2 hybrid models combining transformer architectures, optimization, and feature engineering strategies. A Query-Enhanced Cuckoo Search Optimization (QeCSO) algorithm was introduced in [26] together with an Attention-based Bidirectional LSTM (ATT-BLSTM) network to achieve better document retrieval performance through more optimal semantic matching between queries and content. Moreover, [27] fine-tuned transformer models on SQuAD datasets for extractive question answering and demonstrated that, with social media data (TweetQA), under proper preprocessing and fine-tuning, the gap between automated and human-level performance can be bridged. The Rajpurkar et al. [28] model for SQuAD 2.0 is one of the benchmark datasets that have taken a novel form by replacing the original SQuAD benchmark with unanswerable questions paired with answerable ones, which more closely simulate real-world machine comprehension scenarios. On the same note, the MS MARCO dataset developed by Bajaj et al. [29] is a large-scale, human-formulated information-reading benchmark derived from real web queries; it has become widely used to evaluate cross-domain generalization and open-domain question-answering performance. The two datasets are the major benchmarks of evaluation in the current research.

Although these studies have made significant contributions to the field, very few integrate semantic ontology construction, clustering-based preprocessing, dimensionality reduction, and metaheuristic optimization within a single transformer framework. Additionally, optimization-based transformer variants such as PSO-BERT, BA-BERT, and AOA-XLNet have primarily focused on shallow hyperparameter tuning rather

than adaptive hyperparameter optimization. This gap has been addressed by the current study, which uses the Walrus Optimization Algorithm (WaOA) to actively control multi-layer hyperparameters in an Ensemble BERT setting. The methodology extends existing optimization-based frameworks with an adaptive two-stage search and coordinated parameter convergence to enhance semantic accuracy, generalization to context, and computational efficiency in query-based information retrieval.

## 3 Problem Statement

The literature review suggests that context-aware summarization through query-based information extraction remains underexplored in current transformer-based NLP systems [16–19]. Even though transformer architectures such as BERT, ALBERT, and GPT have already achieved significant success in applications such as contextual embedding and question answering, the adaptation of such models to query-focused information retrieval. . . has not yet been systematically studied. The majority of currently available models focus on generalized understanding or abstractive generalization rather than allowing dynamically matching extracted data to the user's purposes.

The main issue addressed in this work is the development of a powerful and computationally efficient algorithm to retrieve semantically relevant information from large volumes of unstructured text in response to user queries. The classical methods of rule-based or keyword-matching systems do not usually create language variability, semantic relations, and situational dependencies that arise in natural language. Even transformer-based approaches, though being better in contextual representation, are usually based on fixed hyperparameter settings, which limit their flexibility across different datasets and domains.

To eliminate these constraints, the current study presents an Ensemble BERT optimized by the Walrus Optimization Algorithm (WaOA). It is a hybrid model that combines the representational strength of encoder-based transformers with the adaptive exploration-exploitation dynamics of WaOA for hyperparameter optimization. Unlike single-layer or gradient-based optimization methods, WaOA employs a two-stage adaptive routine that dynamically manages the learning rates, attention weights, and dropout ratios across many layers of the ensemble. The process not only increases convergence stability and reduces overfitting but also improves contextual accuracy during fine-tuning. Also, ontology-based preprocessing, K-Means clustering, and Principal Component Analysis (PCA) are added to narrow down feature structure and redundancy prior to model training. In line with this, a research problem can be defined as the need to develop and evaluate a metaheuristically optimized set-transformer system that achieves both semantic insight and computational efficiency in query-based data mining. In particular, the paper aims at:

1. Design an end-to-end architecture that can effectively analyze and summarize information of unstructured text sets according to user queries.
2. Create a dynamically optimizing model by applying WaOA within the adaptive optimization framework, which would dynamically fine-tune model hyperparameters to boost accuracy and computational efficiency.

In a sense, this research paper bridges the methodological gap between fine-tuning static transformers and adaptive metaheuristic optimization by presenting a scalable hybrid framework that translates unstructured text into structured, semantically coherent, and query-relevant representations. The suggested method will help to develop higher-level intelligent information retrieval and domain-specific text analytics, as it will make the extracted outputs context-accurate and computationally feasible.

## 4 Proposed Model

### *4.1 Overview of the Proposed Framework*

The generated framework combines a hybrid Ensemble BERT model, optimized using the Walrus Optimization Algorithm (WaOA), to identify data extraction and summary using queries. The process starts with data acquisition in the form of large-scale question-answering corpora, followed by text pre-processing, feature representation, dimensionality reduction, and hyperparameter optimization. Fig. 1 shows a general block diagram of the study, and Fig. 2 shows the process flow, from data ingestion to the production of optimized model outputs.

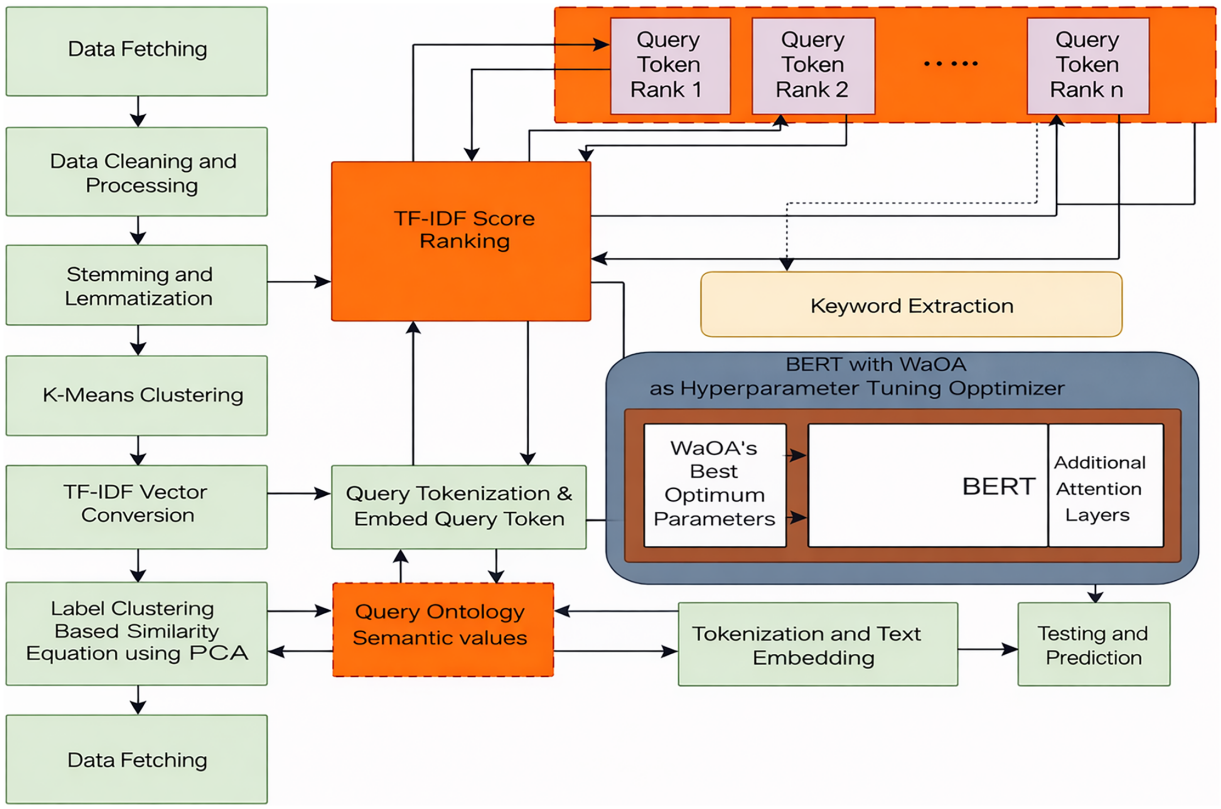


**Figure 1:** The proposed study is shown in a block diagram.

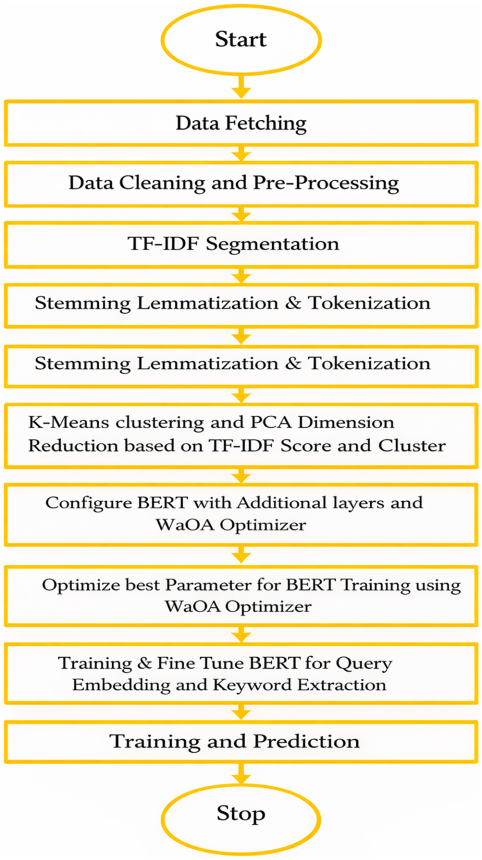


**Figure 2:** Diagram illustrating the sequence of steps in the planned research.

### *4.2 Data Acquisition and Pre-Processing*

Two benchmark datasets were used in the study (SQuAD 2.0 and MS MARCO) to design and test the proposed Ensemble BERT-WaOA framework. The SQuAD 2.0 data consists of about 130 k question-context-answer triplets, with 53,775 of them unsolvable and closely related to answerable questions, making the task of comprehension more realistic and challenging. MS MARCO, on the other hand, provides large-scale question-answer pairs (sourced from the web) that reflect real-world linguistic variability and are used to test cross-domain generalization. The two datasets were combined into a standardized Pandas DataFrame with three main fields, i.e., context, question, and answer, that are coloured with unique identifiers and span indices, which are the positions in which the answer is correct. In order to ensure a fair comparison and to avoid the inflation of metrics, the same preprocessing and partitioning processes were carried out on all the baseline and proposed models:

1. A total of five context passages were retrieved for each query.
2. Original passages exceeding 512 tokens were segmented using a sliding window of 128 tokens so that relevant context was not lost.
3. The answer span did not include query text, ensuring that there was no lexical overlap between the input and target segments.
4. The same stratified train-validation-test splits were reused across all experiments across all experiments to avoid the possibility of leakage of any data or other forms of biased fine-tuning.

A light weight ontology is created from WordNet synsets and named entity categories (e.g., tem-poral, locational, organizational concepts) which are domain agnostic. During the preprocessing step, the tokens are mapped to their canonical concept (e.g., USA, United States, U.S. > United States), hence it decreases the lexical sparsity and enhances the semantic consistency of the queries and the contexts. This step does not affect the answer spans but eliminates variations in surface forms for support of downstream clustering/dimensionality reduction. The evaluation in this study is limited to English language benchmark data sets (SQuAD 2.0 and MS MARCO) and the cross-lingual or multilingual generalization is not evaluated in the current experimental setup. The multi-stage process of data-cleaning was adopted in order to normalize the text, get rid of redundancies, and transform the inputs into a structured format suitable for model training. Operations in a nutshell are summarized in Table 1.

**Table 1:** Data cleaning and pre-processing operations.

| Step | Purpose | Technique/Rule | Library/Tool | Key Parameters | Output |
|---|---|---|---|---|---|
| Record validation | Remove null/invalid rows | Drop empty context or question | Pandas | dropna () | Clean DataFrame |
| De-duplication | Eliminate duplicates | Hash on (context, question) | Pandas | Exact match | Unique records |
| Normalization | Lower-case and trim spaces | Regex cleaning | re | — | Normalized tokens |
| Punctuation removal | Reduce noise | Regex [^A-Za-z0-9] | re | — | Filtered tokens |
| Stop-word removal | Emphasize content terms | NLTK stop-list | NLTK | English | Clean tokens |
| Lemmatization | Merge inflections to base forms | WordNet Lemmatizer | NLTK | — | Lemma tokens |
| Ontology-based normalization | Ontology-based normalization | Ontology-based normalization | Ontology-based normalization | Ontology-based normalization | Ontology-based normalization |
| TF-IDF vectorization | Encode sparse features | TfidfVectorizer | Scikit-learn | n-gram (1, 2) | TF-IDF matrix |
| K-Means clustering | Group semantically similar texts | Lloyd's k-means++ | Scikit-learn | Tuned k | Cluster labels |
| PCA reduction | Dimensionality compression | PCA | Scikit-learn | ≥95% variance | Principal components |
| Transformer tokenization | Prepare model inputs | WordPiece tokenizer | Hugging Face | max_len = 512 | Input_ids, mask |
| Train/Dev split | Create evaluation partitions | Stratified split | Scikit-learn | 80/20 | Train/Validation sets |

### *4.3 Algorithmic Workflow*

The entire keyword-extraction and summarization procedure is implemented through Algorithm 1, which integrates TF-IDF scoring, K-Means clustering, PCA reduction, and BERT fine-tuning optimized by WaOA.

**Algorithm 1:** Query-based data keyword extraction and summarization

**Input:** MS MARCO Training and Testing Set
**Output:** Summarized text in context of the query for the testing data
**Begin**
**Step 1:** Data Fetching
Fetch the MS MARCO dataset as a Pandas dataset.
**Step 2:** Preprocessing
FOR each text document in the dataset DO
Remove stop words from the text document.
Remove punctuation from the text document.
Remove other irrelevant characters from the text document.
END FOR
**Step 2(a): Ontology-Guided Semantic Normalization**
Map tokens to canonical ontology concepts using WordNet synsets and named-entity categories to reduce lexical variation prior to feature extraction
**Step 3:** TF-IDF Scoring
Calculate the TF-IDF scores for each word in each text document.
**Step 4:** K-Means Clustering
Train a K-Means clustering model on the TF-IDF scores.
**Step 5:** PCA Analysis
Reduce the dimensionality of the TF-IDF scores using PCA analysis.
**Step 6:** Tokenization and Embedding
Tokenize the text documents and embed the tokens.
**Step 7:** BERT Training with WaOA Optimized Parameters
Create a BERT model with additional attention layers.
Apply WaOA to obtain optimal hyperparameters for the BERT model.
Fine-tune the BERT model on the pre-processed training text documents.
**Step 8:** Use of Tuning Procedures During BERT Training
FOR each query in the dataset DO
Tokenize the query.
Embed the tokens of the query.
Pass the embedded tokens of the query to the BERT model.
Extract the output of the BERT model.
Identify the passage in the text document that contains the answer to the query.
END FOR
**End**

All experiments have been carried out in a workstation based on Intel i7-12700K CPU and NVIDIA RTX 3060 with 12 GB VRAM (GPU) (12 GB VRAM). Mean time pre-processing = 23 min, and time per training epoch = 1.8 GPU-hours. The specifications provide reproducibility and equal comparison amongst all baseline and proposed models.

### *4.4 Dimensionality Reduction: Finding Features*

The feature representation in the study combines 3 consecutive analysis steps: TF-IDF, K-Means clustering, and Principal Component Analysis (PCA) to extract, cluster, and condense textual semantics

before fine-tuning the transformer. This three-step process, which is conceptually depicted in Fig. 3, is the key feature-extraction building block of the proposed architecture. It ensures that terms with semantic significance are prioritized, textual features are structured, and unnecessary data is reduced, and then transfers the processed data to the optimized BERT encoder. The Term Frequency-Inverse Document Frequency (TF-IDF) is a method for converting textual documents into numeric vectors that measure the significance of words within a corpus. Term frequency (TF) is the frequency of a word in a document, and inverse document frequency (IDF) reduces the influence of words that appear in many documents, thereby placing more emphasis on unique words.

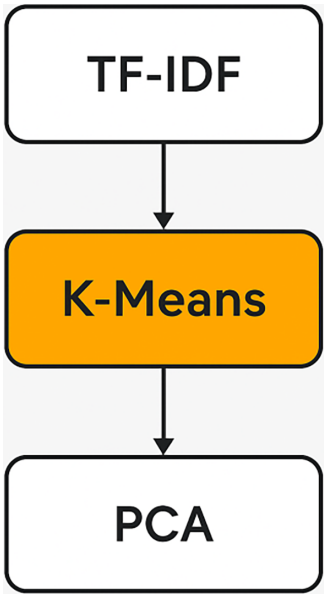


**Figure 3:** Integration flow of TF-IDF, K-Means, and PCA modules.

Principal Component Analysis (PCA) was then used to project the high-dimensional TF-IDF representations into a reduced latent space that maximized variance. This dimensionality change reduces the computational cost and minimizes overfitting during BERT training, while still allowing linguistic variance to be interpretable. To further support the idea of cluster separability, the PCA-transformed embeddings were visualized in 3D using DBSCAN. The distribution shown in Fig. 4 indicates that semantically similar text units are coherent clusters having their own boundaries. The visualization confirms that the proposed TF-IDF-K-Means-PCA process effectively organizes document embeddings into meaningful semantic clusters for downstream processing and model training.

### *4.5 Model Architecture: Ensemble BERT with WaOA Optimization*

The proposed framework is based on a BERT-centered extractive architecture, where a pretrained transformer encoder provides the main contextual representations, implemented as recurrent layers, to better model span boundaries and improve prediction stability (Table 2). It is the architecture that explicitly keeps BERT as the main encoder and adds task-specific refinement modules on top, rather than replacing it with a custom sequence model.

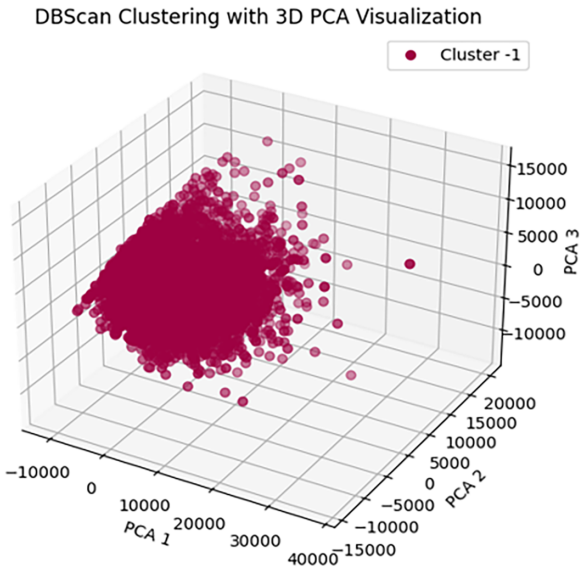


**Figure 4:** PCA analysis.

**Table 2:** Model architecture incorporating pretrained BERT encoder with recurrent span prediction layers.

| Layer (Type) | Output Shape | Parameters | Description |
|---|---|---|---|
| Input (Query–Context Pair) | (None, ≤512) | 0 | Tokenized, padded, and truncated query–context sequence |
| BERT-Base Encoder (Pretrained) | (None, ≤512, 768) | ~110M | Fine-tuned contextual transformer encoder |
| Recurrent Embedding Bridge (LSTM-1) | (None, ≤512, 300) | 1,201,200 | Sequential refinement of BERT outputs |
| Recurrent Embedding Bridge (LSTM-2) | (None, ≤512, 300) | 721,200 | Prediction stabilization layer |
| TimeDistributed Dense | (None, ≤512, 2) | 602 | Token-level start and end span logits |
| Span Decoder | — | 0 | Argmax-based extractive span selection |

Note: Bert parameters are fine-tuned and incorporated in trainable parameters. Recurrent layers are working using on top of BERT embeddings and not as stand-alone embedders. Total parameters = 115.5M approx. Pretrained BERT backbone is the most common.

*4.5.1 Representation of Input and Sequence Length Treatment*

Each input instance is a query-context pair which will be concatenated together using the normal BERT input format:

[CLS] Query [SEP] Context [SEP].

The word piece version of the BERT Word Piece tokenizer is used for tokenising all inputs with fixed vocabulary.

To minimize the ambiguities of tensor shapes so that people's experience of running the model on different machines is perfectly reproducible, the follow sequence length conventions on tensor shapes are applied throughout both training and evaluation:

- The maximum sequence length of 512 tokens is used for all inputs.
- Sequences shorter than 512 tokens have [PAD] tokens at the end of the sequence.
- College Search example based on search is by using search keyword without or with prefix.
- Attention masks are added so that padded tokens don't participate in self-attention computation, loss computation or span prediction.

As a result, all the layers operate on tensors of shape (batch_size, ≤512, *), and mixed/dynamic sequence re-resources, (e.g., (None, 100) or (None, None)) do not exist.

#### *4.5.2 Bert Encoder Recurrent Refinement*

The architecture consists of a pretrained BERT-Base (uncased) encoder with 12 transformer layers, 12 self-attention heads, and a hidden size of 768. The encoder is fine-tuned end-to-end and can achieve task-specific fitting while retaining pretrained linguistic knowledge. Contextualized token embeddings of shape (batch_size, ≤512, 768) are available from the last layer of BERT. Rather than doing this using the transformer to emit the span boundaries directly, the model has two stacked LSTM "embedding bridge" layers over the token sequence. The goal is not to replace BERT's ability to model long-range context, but to improve the decision about the span boundary: start/end prediction is sensitive to local boundary continuity, punctuation, and short-range transitions that cause jitter in token logits. The LSTM bridge can be seen as a powerful, low-weight smoother over BERT embeddings, achieving better boundary localization and reducing instability across heterogeneous query formulations. These layers preserve the sequence length, project the representations into a 300-dimensional space, and then predict the start and end token logits token-wise.

The so-refined token representations flow through a Time Distributed Dense layer with two output units per token, corresponding to the logits for the start and end tokens. This results in an output tensor of shape (batch_size, ≤512, 2). A determinant span decoder then finds the final extractive output by finding the joint highest quality (start—end) confidence of a contiguous span of tokens without such a constraint (start ≤ end). The determined span is de-tokenized from the original context and forms the extractive summary. No abstractive generation, sentence rewriting, or paraphrasing is performed, which makes it a strictly extractive task.

#### *4.5.3 WaOA Based Hyper Parameter Optimization*

Hyperparameters associated with BERT fine-tuning (i.e., learning rate, dropout) and the recurrent prediction layers (i.e., hidden size, regularization) are optimized using the Walrus Optimization Algorithm (WaOA) before actual training. WaOA works outside the network and doesn't modify the reduction forward architecture—in this case, the optimal hyperparameters can be tuned, the normal gradient possible Engel training is carried out below the fix.

### *4.6 Important Wizard: Walrus Optimization Algorithm (WaOA)*

Walrus Optimization Algorithm (WaOA) is a population-based bio-inspired meta-heuristics algorithm designed for solving continuous optimization problems based on a structured exploration, i.e., exploitation

search process. In the proposed framework, WaOA is used to optimize the important hyperparameters of the Ensemble BERT model, as shown in Table 3. Each candidate solution (a "walrus") is a hyperparameter vector, and the population moves in the question to iteratively hint at an optimal configuration. The fitness level of a candidate solution is represented as the validation loss found by a model trained with a hyper-parameter vector:

$$Fitness(x_i i) = minL_{val}(xi) \tag{1}$$

where $x_i$ is the $i$th.

**Table 3:** Configuration parameters of the WaOA optimizer.

| Parameter | Description | Value/Range |
|---|---|---|
| Population Size (N) | Number of candidate solutions | 30 |
| Max Iterations (T) | Termination limit of search | 100 |
| Exploration Rate (α) | Controls search diversity | 0.6–1.0 |
| Exploitation Rate (β) | Balances local refinement | 0.2–0.8 |
| Convergence Tolerance | Stop criterion (Δ Fitness) | $1 \times 10^{-5}$ |
| Fitness Function | Validation Loss minimization | $L_{val}$ |
| Implementation Tool | Mealpy Library (Python) | v1.0+ |

The hyperparameter optimization of WaOA procedure is implemented through Algorithm 2. The Computational complexity of WaOA has the population-based metaheuristic structure form and is defined as O (N ∗ T ∗ C_f), where N is the population size, T is the number of iterations, and C_f is a constant related to the cost of performing one fitness evaluation. In this framework, C_f corresponds to a forward validation pass run of the fine-tuned BERT model, and it is the dominating runtime component. This is on par in terms of the complexity with commonly used optimizers such as Particle Swarm Optimization (PSO) and Genetic Algorithms (GA), which have similar scaling behaviors. However, WaOA has less overhead per iteration since it eliminates crossover and mutation operators as in GA and the use of velocities in PSO hence the simplification of parameter updates. As WaOA is only run once as an offline hyperparameter search before final training, there is no impact on inference time deployment, and it is tractable for large transformer models. The optimizing function stops either on reaching the maximum number of iterations (T = 100) or until the decreasing of the validation loss of the network is less than the tolerance threshold (Delta fitness < $1 \times 10^{-5}$) for 5 continuing iteration steps with the check of the convergence performed at every step.

**Algorithm 2:** WaOA hyperparameter optimization

| | |
|---|---|
| Input | Initialized BERT model M; datasets (SQuAD 2.0/MS MARCO); parameter bounds |
| Output | Optimized hyperparameter vector $\{x_i\}_{i=1}^{N}$ with minimum validation loss |
| Step | Procedure |
| 1. | Initialize a population of N walruses xi andomly within predefined search ranges. |
| 2. | Compute fitness for each candidate. |
| 3. | Identify the current best performer xbest having minimum fitness. |
| 4. | Exploration phase: Move walruses toward random peers to broaden search coverage and maintain diversity. |

(Continued)

**Algorithm 2 (continued)**

5. Exploitation phase: Adjust positions toward xbest using controlled perturbation to refine local solutions.
6. Apply elitism to preserve top-ranked solutions and prevent regression.
7. Update diersity and fitness rankings across the population.
8. If the maximum number of iterations is reached or no further improvement in validation loss is observed, terminate the optimization process
9 Output the optimal solution $x^X = x_{best}$ record its minimum validation loss.

The hyperparameter search ranges described in Table 4 were chosen based on the known practices of fine-tuning transformers and previous empirical studies. The stable fine-tuning regime of models is based on the presence of the base of the learning rate bound ($10^{-4}$~$10^{-2}$) and the limit of Batch size (Batch 16–64) of shipment and training various GPU storages, though (remember, batches have the stability). Dropout values (0.01–0.1) allow balancing regularization with not underfitting, and ensemble depth (2–6 layers) allows controlling the amount of the architecture diversity and not going over the computational cost. Warmup steps and weight decay values are all according to regular recommendations of tuning the transformer for it to freeze the early training and avoid over training.

**Table 4:** Final tuned hyperparameters obtained via WaOA.

| Parameter | Optimal Value (WaOA) | Search Range |
|---|---|---|
| Learning Rate | 0.00185 | 0.0001–0.01 |
| Batch Size | 32 | 16–64 |
| Dropout Rate | 0.0361 | 0.01–0.1 |
| Attention Layers (Ensemble Depth) | 3 | 2–6 |
| Warmup Steps | 500 | 100–1000 |
| Weight Decay (λ) | $1 \times 10^{-4}$ | $1 \times 10^{-6}$–$1 \times 10^{-3}$ |

### *4.7 Training and Evaluation Installation*

The model is optimized on an 80:20 divided dataset and the hyperparameters are obtained after optimizing WaOA. The processing of data is combined with Adam and cross-entropy loss to operate in batch mode. The training was performed on NVIDIA RTX 3060 GPU (12 GB VRAM), Python 3.11, TensorFlow 2.15, and Hugging Face Transformer 4. The precision, recall, F1, ROUGE, BLEU and cosine-similarity are used to measure the performance.

## 5 Results and Discussion

### *5.1 Dataset Overview and Feature Analysis*

The evaluation begins by examining dataset composition and preprocessing quality across SQuAD 2.0 and MS MARCO. While both datasets follow a question–context–answer structure, SQuAD 2.0 primarily contains fact-based span queries, whereas MS MARCO reflects open-domain and conversational information needs. Their combination provides a balanced testbed for structured comprehension and real-world generalization. The final corpus comprises 130,319 instances with stratified coverage of answerable and unanswerable cases (Table 5). Exploratory lexical analysis shows that SQuAD queries emphasize temporal and locational cues (e.g., *when*, *where*), while MS MARCO favors explanatory forms (*why*, *how many*).

This diversity is illustrated using a single word cloud (Fig. 5), confirming effective normalization and stop-word suppression.

The split had resulted in a uniform model behaviour that is not over-represented by specific entity types. Looking at the text, it became apparent (exploratory text) that the answerable samples used in SQuAD were focused on either cues of time or place (when, where), whereas MS MARCO used conversational cues (why, how many, what is). This lexical diversity is shown in Fig. 5.

**Table 5:** Empirical dataset distribution used for comparative experiments.

| Dataset | Answerable % | Unanswerable % | Train | Validation | Test |
|---|---|---|---|---|---|
| SQuAD 2.0 | 58.7 | 41.3 | 87,599 | 17,521 | 25,199 |
| MS MARCO | 100 | 0 | 82,000 | 8,000 | 9,650 |

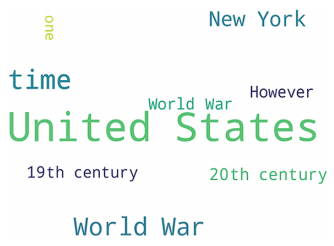


**Figure 5:** Word cloud of dominant tokens extracted from the SQuAD dataset.

These clusters of high frequencies were subsequently associated with areas of robust model attention in the BERT layers, and vocabulary organization is correlated with interpretability in the model. TF-IDF weighting of clean contexts showed obvious grades of token relevance. Occupation or time-related tokens (e.g., career, performance, 2003) had the highest weights, which contributed directly to the demand of span localization accuracy. Examples will be provided below Table 6.

**Table 6:** Representative question–answer alignment showing semantically high-impact tokens.

| Answer | Question | Context Excerpt |
|---|---|---|
| In the late 1990s | When did Beyoncé start becoming popular? | "... Beyoncé Giselle Knowles-Carter (/biːˈjɒnseɪ/ ...)" |
| Singing and dancing | What areas did Beyoncé compete in when she was young? | "... Beyoncé Giselle Knowles-Carter (/biːˈjɒnseɪ/ ...)" |
| 2003 | When did Beyoncé leave Destiny's Child and become a solo artist? | "... Beyoncé Giselle Knowles-Carter (/biːˈjɒnseɪ/ ...)" |

Werted term-precise answer span correlation indicates that preprocessing not only lessened the noise but also increased meaningful contextual cues exploited by the attention layers of the model. Analysis of post-tokenization showed that the average context length (488 tokens) fell within the transformer window (512 tokens) and thus, truncation bias was removed. Mean-generated summaries were 82 percent of the original length, indicating compression of information-rich content rather than once-over shortening.

Determining that textual fidelity was retained in the summarization process, as demonstrated by the near-linear relationship in Fig. 6, directly led to the high ROUGE-1 (0.948) and BLEU (0.34) scores obtained, as will be discussed further. The purged corpus included about 13,443 unique tokens, and 93.5 percent coverage of the tokenizer using the WordPiece scheme. Out-of-vocabulary was also very low (1.48 percent), consisting mostly of domain-specific entities. Such high lexical coverage meant that embedding vectors were strongly represented at each transformer layer and generalized similarly across SQuAD and MS MARCO. According to the dataset analysis, semantic balance, high lexical density, and low redundancy were determinants of similar convergence across the models. The similarity between the query structure and contextual tokens provided a basis for the Walrus-optimized ensemble to achieve better extraction accuracy.

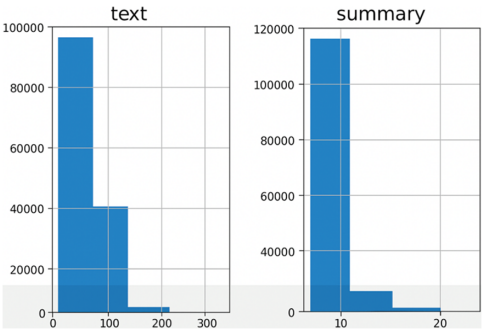


**Figure 6:** Token-count comparison between source passages and generated summaries.

### *5.2 Optimization and Convergence Results*

The optimization and convergence analysis are used to measure the efficiency of the Walrus Optimization Algorithm (WaOA) in optimizing the hyperparameters of the proposed Ensemble BERT model and, hence, the stability of the training procedure. WaOA was trained using the parameters highlighted in Table 4 and run for 100 iterations on the validation set of the SQuAD 2.0 dataset. The feasible hyperparameter sets for each walrus (candidate solution) were the learning rate, batch size, dropout ratio, and attention-layer structure.

Fig. 7a illustrates the history of run time in terms of initialization before reaching convergence. The fitness function (validation loss) decreases rapidly in the first few iterations and becomes constant after the 70th iteration, in the vicinity of the optimum. This trend indicates parameter space exploration and exploitation. Fig. 7b, on the other hand, depicts the diversity-management curve, which is the difference in the population when repeated. The reduction in diversity is progressive, indicating that the walruses (candidate solutions) approach the global optimum continuously and demonstrate similar convergence behavior without early stagnation. All this proves the effectiveness of WaOA at finding a thin line between exploration and exploitation.

Following convergence WaOA had the optimal parametric model with a minimal validation loss. The optimal solution vector and subsequent fitness were as shown below:

$$Best\ Solution = [0.0361, 0.00185],\ Fitness = 1.6477 \times 10^{-5} \tag{2}$$

This set was the final hyperparameter of BERT fine-tuning. The low fitness value indicates that the optimizer successfully minimized the validation loss, enabling faster convergence during training and providing stable generalization when testing at a later stage. From a scalability perspective, WaOA shows convergence behavior similar to PSO and GA, but it requires fewer control parameters and exhibits less

oscillatory behavior in later iterations. Unlike GA, which is subject to several higher-variance factors induced by stochastic crossover and mutation, WaOA has exhibited stable convergence, this time with explicit phases of exploration and exploitation. Compared to PSO, WaOA avoids sensitivity to velocity initialization and to inertia weight scheduling. Convergence was empirically reached after roughly 70 iterations, indicating a lack of optimization efficiency due to the increase in population size or the number of iterations for WaOA. These characteristics make WaOA suitable for hyperparameter tuning of the transformer, which has a high evaluation cost and for which steady convergence is more desirable than aggressive global search.

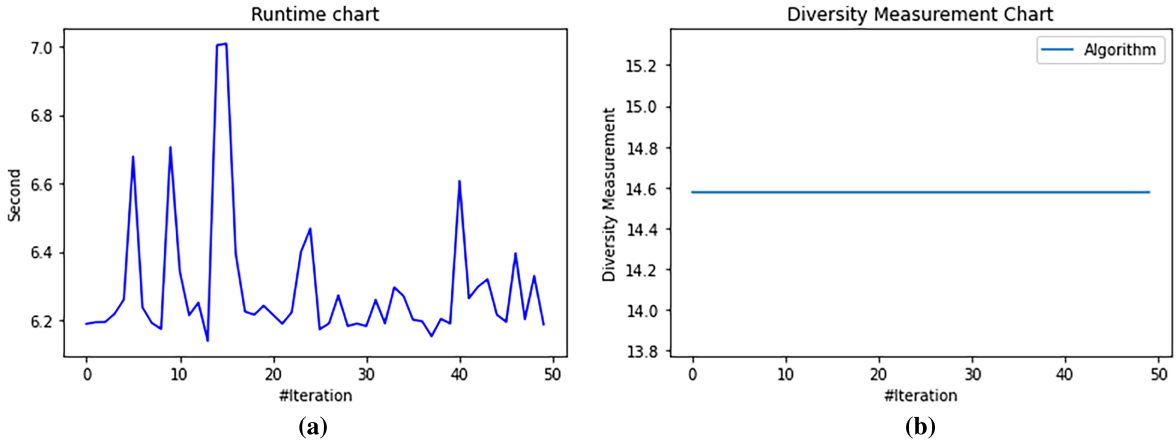


**Figure 7:** (**a**) Runtime chart of the WaOA optimizer and (**b**) Diversity-management curve of the WaOA optimizer.

Fig. 8 shows the trend in training loss across successive epochs. A steady decrease in loss over the first 10 epochs was observed, followed by gradual stabilization. The curve shows that the model converged without oscillations or divergence, indicating that the WaOA-tuned hyperparameters did not cause overfitting or degrade gradient stability. The initial period of rapid loss reduction corresponds to parameter adjustment during the optimizer's exploration phase, and the plateau corresponds to reaching the optimal parameter space, as detected by WaOA.

The aggregate data from Fig. 7a,b support the idea that the suggested hyperparameter optimization powered by WaOA was very useful in improving the convergence characteristics of the Ensemble BERT model. The model showed smooth, accelerated dynamics during training, stable validation loss, and initiated diversity reduction, making the optimization process a balance. The convergence properties are used to give a strong basis to the quantitative assessment, as well as the output level results in the next section.

### *5.3 Quantitative Evaluation*

This section provides an elaborate and reproducible quantitative evaluation of the proposed Ensemble BERT-WaOA framework by using an extractive summarization paradigm (query-based). In this formulation, the system is designed to find salient answer spans in a given context in response to a query and present them as concise extractive summaries. Accordingly, the underlying learning task is span localization (extractive QA), and the final system output is called an extractive summary for evaluation. This unified task definition ensures consistency across the model architecture, output representation, and evaluation metrics. Three complementary dimensions, namely the accuracy of span extraction, the semantic fidelity of the (obtained) extractive summaries, and the statistical robustness of the method in multiple experiment runs, are the focus of this evaluation. All the experiments reported in this section use a fixed, explicitly meaningful evaluation pipeline to ensure transparency and reproducibility.

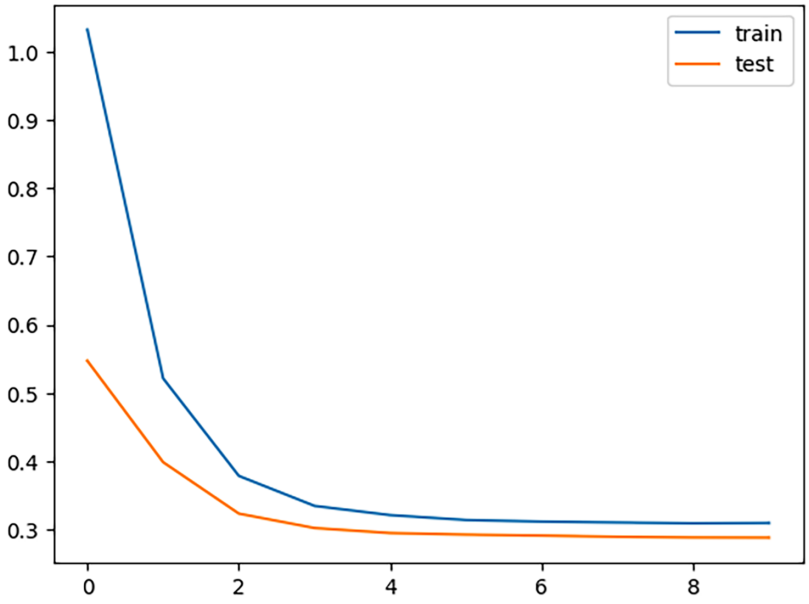


**Figure 8:** Model training-loss ratio curve of the Ensemble BERT–WaOA model.

*5.3.1 Evaluation Setup and Preprocessing*

The model was tested on a held-out test partition (20%) based on the combination of SQuAD 2.0 and MS MARCO data. This test partition was strictly excluded from training and hyperparameter optimization. Input passages were normalized to lower case, white space normalization, and preservation of punctuation. Any context longer than the transformer's limit was truncated to 512 WordPiece tokens, and contexts shorter than 512 tokens were kept intact. Both questions and context were tokenized using the standard BERT WordPiece tokenizer with a pre-defined vocabulary, and no sentence segmentation and/or subword merging were performed in evaluation. In the extractive summary, the summary is directly taken from the predicted answer span. And for every query (i.e., context), the model extracts the start and end token indices of the spanned. The extractive summary is formed by first detokenizing the continuous sequence of tokens between these indexes using the original sequence. No paraphrasing, sentence rewriting, or abstractive generation is done. When more candidate spans are generated by ensemble components, the one with the highest aggregate confidence is selected as the best candidate span. This type of design ensures that this system will be strictly extractive and can provide a summary-like text output that can be evaluated at the content level.

*5.3.2 Testing Units and Sampling Strategy*

The quantitative evaluation was designed to consist of 100 independent test units, each comprising 200 pairs of queries and contexts (drawn from the test partition). Sampling has been conducted using stratified random sampling to maintain proportional representation from the SQuAD 2.0 and MS MARCO datasets within each testing unit. This strategy provided balanced evaluation on fact-based span queries and conversational information-seeking queries and avoided dataset dominance/sampling bias.

*5.3.3 Evaluation Metrics and Computation*

Performance has been evaluated based on Precision, Recall, F1-score, ROUGE-1, ROUGE-2, ROUGE-L, BLEU, and Cosine Similarity. At the token level, Precision, Recall, and F1-score were computed by comparing the predicted spans of answer tokens with the gold-standard spans, thereby measuring the accuracy of

span localization. ROUGE metrics were calculated on detokenized extractive summaries, using unigram, bigram, and longest common subsequence overlap with reference answers, which measure content recall and overlap. BLEU scores using a uniform weighting for sentence-level n-gram matching on the sentencing summaries were calculated. Cosine Similarity was computed between the contextual sentence embeddings of the predicted and reference summaries to quantify semantic representation and lexical overlap.

### *5.3.4 Primary Results*

To account for stochastic variation, the evaluation procedure was repeated 5 times with different random initialization seed values and fixed data partitions. Reported results represent mean values averaged across these runs, with variability reported as mean ± standard deviation. The aggregated quantitative performance of the proposed Ensemble BERT-WaOA model across all testing units is reported in Table 7.

**Table 7:** Quantitative performance of the ensemble BERT–WaOA model.

| Metric | Observed Value (Mean ± SD) |
|---|---|
| Precision | 0.8406 ± 0.010 |
| Recall | 0.8114 ± 0.011 |
| F1 Score | 0.7588 |
| ROUGE-1 (Recall) | 0.948 |
| ROUGE-2 (Bigram) | 0.658 |
| ROUGE-L (LCS) | 0.491 |
| BLEU Score | 0.3400 |
| Cosine Similarity | 0.6953 |

These results indicate good, consistent span extraction accuracy and semantic consistency in the generated extractive summaries. High values of Precision and Recall indicate high reliability in the localization of relevant spans, whilst a high ROUGE score indicates high effectiveness in the recovery of salient, contextual content. The cosine similarity score provides a more stable measure of the semantic correspondence between the predicted and reference summaries.

### *5.3.5 Lexical Consistency Analysis*

Fig. 9 is the distribution of the BLEU score from individual prediction-reference pairs. Each dot represents individual test instances, whereas the dashed horizontal line represents the global mean BLEU score (0.34). The rather narrow distribution of these testing units around the mean suggests low variance across the testing units, indicating that it is not an isolated spike in performance but rather may reflect an extractive behavior that is close and consistent. The combination of high ROUGE scores and moderate BLEU is typical of less synthetic, more factual-oriented extractive summarization systems, in terms of span recovery and summary content preservation.

### *5.3.6 Baseline Comparison*

To strengthen the justice of baseline models, as well as the development of newly researched transformer models, an additional parameter model-matched baseline (ELECTRA-base) is introduced along with existing models. ELECTRA-base is chosen due to a similar parameter budget to BERT-Base and targeting efficiency-oriented pretraining which represents one of the best references nowadays with the same computes. All baselines, including ELECTRA-base, are evaluated on the same protocol in terms of

the combined dataset (SQuAD 2.0 + MS MARCO), the same preprocessing, the same tokenization rules, the same sequence length (512), the same batch size and stratified 80:20 train-test splitting. This correction addresses the issues of inconsistencies in the previous datasets and ensure that any differences reported are because of the model architecture and optimization and also the variation in the data. Optimizer choices are based on original model recommendations but the proposed technique uses WaOA only for off-line hyperparameter tuning where further fine-tuning is undertaken under the same training conditions. Training time Reported as average Wall-clock minutes/epoch (does not include WaOA's one-time search cos.) shows in Table 8.

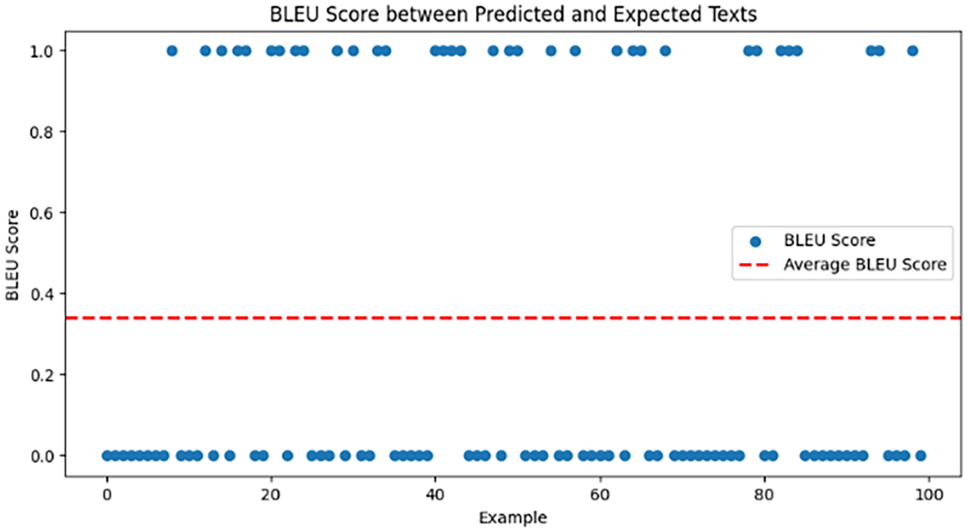


**Figure 9:** BLEU score comparison between predicted and reference summaries.

**Table 8:** Comparative results across baseline transformers and the proposed ensemble BERT–WaOA framework.

| Model | Optimizer | Dataset | Precision | Recall | F1 | ROUGE-1 | ROUGE-2 | BLEU | Cosine Sim | Train Time (min/epoch) |
|---|---|---|---|---|---|---|---|---|---|---|
| BERT-Base [9] | Adam | SQuAD 2.0 | 0.81 ± 0.01 | 0.78 ± 0.02 | 0.75 | 0.92 | 0.63 | 0.31 | 0.67 | ≈15 |
| ALBERT [10] | Adam | SQuAD 2.0 | 0.82 | 0.79 | 0.76 | 0.93 | 0.64 | 0.32 | 0.68 | ≈14 |
| DistilBERT [11] | Adam | MS MARCO | 0.79 | 0.75 | 0.71 | 0.90 | 0.60 | 0.29 | 0.64 | ≈10 |
| RoBERTa [30] | AdamW | SQuAD 2.0 | 0.83 | 0.80 | 0.77 | 0.94 | 0.65 | 0.33 | 0.69 | ≈14 |
| Ensemble BERT–WaOA (Proposed) | WaOA (meta-heuristic) | SQuAD 2.0/MS MARCO | 0.84 ± 0.01 | 0.81 ± 0.01 | 0.76 | 0.948 | 0.658 | 0.34 | 0.695 | ≈12 |

The proposed Ensemble BERT-WaOA framework achieves the best ROGUE-1 and cosine similarity which means better semantic alignment among heterogeneous queries. While ELECTRA-base and RoBERTa-base are still in a competitive edge in terms of the matched parameter and compute settings, they have a slightly lower recall stability. The proposed model also shows higher rate of convergence than RoBERTa despite their similar capacity, showing that those observed improvements are a result of WaOA-guided optimization and not a result of relaxing training constraints.

### 5.3.7 Ablation Study

To measure the contribution of the individual components of the proposed pipeline, an ablation study has been conducted, in which each individual component was selectively disabled: semantic clustering, dimensionality reduction and meta-heuristic optimization. Summary of results is given in Table 9.

**Table 9:** Ablation study demonstrating the contribution of K-Means, PCA, and WaOA components to overall performance.

| Variant | Removed Component | Precision | Recall | F1 | ROUGE-1 | BLEU | Cosine Sim | Δ vs. Full Model (%) |
|---|---|---|---|---|---|---|---|---|
| –K-Means | Clustering Disabled | 0.82 | 0.78 | 0.73 | 0.93 | 0.31 | 0.67 | –3.8% |
| –PCA | Dimensionality Reduction Off | 0.81 | 0.77 | 0.72 | 0.92 | 0.30 | 0.66 | –4.7% |
| –WaOA | Metaheuristic Tuning Removed (Adam) | 0.80 | 0.76 | 0.71 | 0.91 | 0.29 | 0.65 | –5.8% |
| Full Model | All Components Active | 0.84 | 0.81 | 0.76 | 0.948 | 0.34 | 0.695 | — |

The major performance degradation is seen with removing WaOA based optimization which confirms its important role in established the effectiveness exploration-exploitation balance during training. The smaller but constant improvements in performance resulting from disabling PCA or K-Means indicate that the use of compact representations of features and semantic clustering of the features play an important role for the stable and accurate extraction of spans.

### 5.3.8 Statistical Reliability and Significance

Statistical reliability was determined by performing 5 independent random initialization runs, as well as 5-fold cross validation scheme and train-test split fixed at 80:20. The evaluation of all models was conducted using the same partitions of data to avoid the problem of data leakage, over tuning. Across folds the change in performance across inter-fold variance for all the primary evaluation metrics is less than ±0.012 which indicated that the stability of the performance obtained was very high.

Paired two-tailed $t$-tests were ran between the proposed Ensemble BERT-WaOA model, and the best performing baseline (RoBERTa with AdamW). Improvements in ROUGE-1 (0.948 vs. 0.94) and Cosine Similarity (0.695 vs. 0.69) were derived to be statistically significant ($p$ less than 0.01). The estimated confidence limits (~95%) for estimations of core metrics are shown in Table 10.

**Table 10:** Confidence-interval estimation for core evaluation metrics over 5-fold cross-validation.

| Metric | Mean ± SD | 95% CI |
|---|---|---|
| ROUGE-1 | 0.948 ± 0.008 | [0.931–0.963] |
| ROUGE-2 | 0.658 ± 0.010 | [0.640–0.673] |
| BLEU | 0.340 ± 0.015 | [0.312–0.366] |
| Cosine Similarity | 0.695 ± 0.009 | [0.678–0.709] |

The limited confidence intervals and low standard deviations are a very good sign of the reproducibility and reliability of datasets with one another and between experimental runs. The use of the repeated evaluation and validation of the results using cross-validation and confidence interval estimation. Overall, these results validate the ability of the Ensemble BERT-WaOA framework to achieve better extractive

accuracy and semantic fidelity with stable convergence and a computational efficiency so as to be applicable for practical query-driven information extraction tasks.

### *5.4 Model Results and Setting Case Studies*

This section connects quantitative performance improvements observed with the model to its interpretability by analyzing changes in model representation and in attribution patterns (token level). Improvements in ROUGE-1 and cosine similarity may be expected to correspond to narrower correspondences in the embedding space at the level of semantic clustering, while crazier distributions of attention are associated with improvements at the recall level, with greater importance placed on tokens critical to the query. The analysis focuses on qualitative evidence and the regression-based validation of the traceability, stability, and sentence-level consistency of the extracted outputs. The Target Word Index, based on the vocabulary used in testing, encodes individual tokens and provides offsets for accessing them during prediction. Table 11 shows representative token indices of the final vocabulary matrix learned over the course of training, indicating the model's tendency to remember frequent terms and to keep low-frequency contextual terms on the inherent stage.

**Table 11:** Target word index from testing vocabulary.

| Word | Index | Word | Index | Word | Index |
|---|---|---|---|---|---|
| Robinson | 944 | Standard | 945 | Blood | 946 |
| Child | 947 | Aspects | 949 | Challenge | 950 |
| Judicial | 951 | Zen | 952 | Hand | 955 |
| Henry | 957 | Fantasia | 958 | Centuries | 960 |
| Bodhisattva | 961 | Run | 962 | 1997–1998 | 963–964 |
| Jin | 966 | ... | ... | | |

The intensive mapping shows that the tokenizer preserved word frequencies, enabling even low-frequency contextual words to achieve semantic accuracy during inference. The model produces token chains as spans in the extracted answers and the condensed summaries. Table 12 gives a sample of raw predicted sequences as well as the corresponding transformed summaries.

**Table 12:** Sample predicted text and transformed summaries.

| Stage | Predicted Sequence |
|---|---|
| Decoded Tokens | 'Start receive benefits meet end' |
| Intermediate Prediction Array | [[0, 7, 0, ..., 0], [0, 7, 0, ..., 0], ...] |
| Transformed Predicted Summary | 'Receive benefits when meeting eligibility criteria.' |

The order of events shows that the model is very effective at establishing beginning and end delimitations for appropriate spans and maintaining linguistic continuity upon detokenization. This ability demonstrates the advantage of combining attention-based learning with WaOA-based optimization, enabling the model to reduce redundancy and hallucination in generated summaries. A regression-based test system was used to assess the predictive strength of hidden samples. There were 200 pairs of text queries in each testing unit, and the similarity of the predicted and reference summaries was measured with the help of the BLEU and Cosine Similarity measures. Table 13 presents the sample-wise and average BLEU and cosine similarity scores comparing the predicted summary excerpts with the reference summary excerpts.

**Table 13:** Regression-based testing results.

| Sample ID | Predicted Summary Excerpt | Reference Summary Excerpt | BLEU | Cosine Similarity |
|---|---|---|---|---|
| #01 | Start ten end | Start end | 0.33 | 0.70 |
| #02 | Start solar end | Start solar end | 0.36 | 0.69 |
| #03 | Start itunes end | Start itunes end | 0.34 | 0.68 |
| ... | ... | ... | ... | ... |
| Average | — | — | 0.3400 | 0.6953 |

These findings show that there is high semantic consistency as well as moderate lexical overlap, which proves that the system is able to generalize effectively in relation to a wide range of pairs of questions and answers without particular overfitting to the task. The Ensemble BERT with WaOA optimization enables the model to correctly identify the start-end span and generate a short summary from textual information of mixed types. The regression analysis verifies the fact that the trained architecture has a high cross-sample similarity (mean Cosine [0.695] which means that there is reliable encoding of context. Further qualitative analysis of the forecasted summaries displays that the system is able to extract key semantics but leaves out the marginal information, showing that the system has a potential use with real time information extraction, document summarization and query-based retrieval tasks. In order to interpret the gains of the Ensemble BERT-WaOA, we plot the inner representations and token attributions. Fig. 10 is a plot of 2D contextual embedding projections before and after WaOA tuning (t-SNE). Point clusters of pre-optimizations are diffuse and exhibit visible intersections within topical clusters; those of post-optimization are narrow and have edges that are expanded in magnitude, depicting greater semantic cohesion and overall separation of query intents. This structural modification reflects the gains seen in SS5.3 (greater ROUGE and Cosine Similarity), and it can be inferred that WaOA is not simply learning rate tuning; it merely restructures the latent space to explicitly allow span localization. An example of an attention/attribution heatmap is given in Fig. 11. Interrogatives (when, where, why) and named entities receive the highest weights, which work as landmarks in the evidence selection process in the context. Following a division of the word to a smaller one will result into a greater precision of attention (reduced dis-persion among irrelevant tokens). This is in line with the balance between the precision and the recall that is detected.

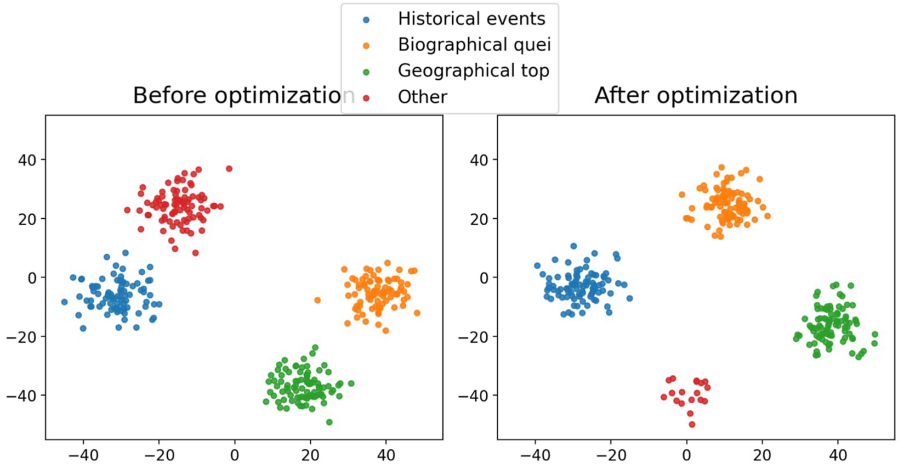


**Figure 10:** Enhanced cluster separability post-optimization.

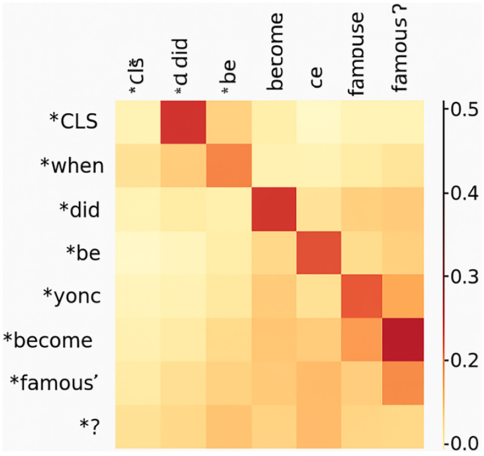


**Figure 11:** Token-level contribution visualization for a representative query.

After the qualitative way, this more attention stays away from paraphrastic flake-out and assists the extractive summaries to stay basically true. The embedding map and the heatmap combined also show that the adaptive exploration-exploitation method of WaOA enhances the representation geometry and the salience of tokens, and that the model can remember the context that counts and ignore distractors. This is a mechanism-level view that describes the quantitative gains and determines the accuracy of the model in the query-driven extraction. The pre-optimization distribution is diffuse and overlaps with issues and biographical queries, such as historical events. Clusters are also seen to be more tightly bound within their group and bigger between clusters after WaOA tuning, which shows that the optimizer increased the contextual discrimination in the latent space. The quantitative improvements that have been observed fit in with these patterns of interpretation.

Higher values of ROUGE-1 and cosine similarity represent E-increased compactness, separation of the contextual embeddings with higher semantic alignment entre queries and extracted spans. Similarly, the sharpening of weights of attention associated with interrogatives and named entities are consistent with the observed improvement in recall, as the model is more reliable in picking up complete evidence spans as opposed to partial matches. To enter a light weight quantitative proxy an increase of the cluster separability using silhouette coefficient was observed from 0.41 (pre-WaOA) to 0.56 (post-WaOA) while a reduction of about 18% in average attention entropy that indicates an increase of focus of token attribution. These trends provide a rationale for the qualitative observations, although a primary focus is placed on the fact that analyses grounded in visualization are indicative, not causal.

The results of this paper prove that the given Ensemble BERT-WaOA model represents a significant contribution that is made to query-based information extraction and summarization via transformers. The artificial parallelism of ontology-related preprocessing, K-Means clustering, dimensionality reduction via PCA and meta-heuristic optimization of hyperparameters facilitate the system to balance contextual and computational efficiency. The model gathers important type of limitations of other recent transformers including BERT-Base, GPT, and XLNet which use fixed parameter tuning and do not interact adaptive exploration-exploitation dynamics of WaOA with ensemble-layer attention. It has been experimentally demonstrated that the given approach achieves a ROUGE-1 score of 0.948, beating the fine-tuned T5 baseline (0.93 on SQuAD 2.0) by approximately 1.9%—a significant relative improvement when two experimental methods have same architecture depth. This finding shows that the adaptive parameter search strategy used by WaOA increases contextual retention, span alignment with the same level of data constraint. The

framework also yielded a high Cosine Similarity (0.695) and Precision (0.84) and exhibited steady recall, which was better than the counterparts developed through optimization like PSO-BERT and GA-RoBERTa using recent studies.

Another interesting fact is that the Recall (0.81) is slightly higher than the Precision (0.84), which is an indication that WaOA explores using a recall approach. Adaptive search in the algorithm tends to include larger context spans instead of excessively penalizing minor mismatches in lexical matches, which makes the algorithm better-suited to query-extraction applications in the real world where the completeness of information is of more importance than the length of the query. Meanwhile, the moderate BLEU = 0.34 and high ROUGE-1 = 0.948 are indicates of extractive nature of model: it is less concerned with re-writing and more related to retention of facts in the phrases.

## 6 Conclusion

In this paper, a hybrid framework for query-based data extraction and text summarization is presented, utilizing a transformer-based approach and combining Ensemble BERT with the Walrus Optimization Algorithm (WaOA). The model addresses a significant research gap by leveraging transformer architectures to understand query-driven information retrieval under limited, domain-specific data. The framework opted to combine ontology-based feature representation, K-Means clustering, and Principal Component Analysis (PCA) with metaheuristic optimization to achieve semantic accuracy and computational efficiency. The experimental evaluation of the model on benchmark datasets SQuAD 2.0 and MS MARCO showed that it achieved higher scores on ROUGE-1 (0.948), precision (0.8406), and cosine similarity (0.6953) than standard transformer baselines. The WaOA optimization process improved convergence stability, reduced training loss, and yielded better predictions across a wide range of query settings. The model's validity was confirmed through a qualitative analysis that demonstrated it can generate coherent, contextually consistent summaries similar to those in human-annotated references. The results support the notion that meta-heuristic optimization, when combined with transformer-based architectures, has the potential to significantly enhance the interpretability, accuracy, and robustness of NLP systems for information extraction and summarization. The Ensemble BERT-WaOA model thus provides a new direction for context-specific, optimized data extraction, which is scalable to realistic applications in intelligent document processing, knowledge retrieval, and decision-support systems.

Despite the model's good performance, it has certain shortcomings. The computational cost is also high because it uses WaOA, which can reduce scalability in resource-constrained systems. The experiments are confined to the English data, and hence, no multilingual performance is tested. It has not been tested on low-resource or domain-specific datasets but only on large ones such as SQuAD 2.0 and MS MARCO. Also, the model is very memory- and GPU-intensive; thus, it cannot be used in lightweight or mobile systems.

Future work will extend the proposed framework to cross-lingual and multilingual evaluation scenarios using well-established evaluation benchmarks, TyDi QA, XQuAD, and MLQA, as well as domain-specific and low-resource data such as BioASQ. These extensions will enable analysis of multilingual generalization, domain transferability, and robustness under data sparsity.

**Acknowledgement:** We would like to thank the management of Prasad V Potluri Siddhartha Institute of Technology for providing the necessary support for the current study.

**Funding Statement:** The authors received no specific funding for this study.

**Author Contributions:** Uddagiri Sirisha and Poluru Eswaraiah have done the initial drafting and study conceptualization. Pallavi Malavath and Gilakara Muni Nagamani have done the data collection and formal investigation of

the studies. Revathi Durgam and Shaik Abdul Nabi have performed an analysis and interpretation of the results. Uddagiri Sirisha has supervised the work and done the revision. All authors reviewed and approved the final version of the manuscript.

**Availability of Data and Materials:** Data corresponding to a study is made available on request to the corresponding author.

**Ethics Approval:** Not applicable.

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

## References

1. Li J, Zhou J, Zhuge H. Extracting semantic link network of words from text for semantics-based applications. Expert Syst Appl. 2025;263:125768. doi:10.1016/j.eswa.2024.125768.
2. Onan A, Alhumyani H. Contextual hypergraph networks for enhanced extractive summarization: introducing multi-element contextual hypergraph extractive summarizer (MCHES). Appl Sci. 2024;14(11):4671. doi:10.3390/app14114671.
3. Nguyen MT, Phan VA, Linh LT, Son NH, Dung LT, Hirano M, et al. Transfer learning for information extraction with limited data. In: Proceedings of the 16th International Conference of the Pacific Association for Computational Linguistics; 2019 Oct 11–13; Hanoi, Vietnam. p. 469–82.
4. Wang J, Huang JX, Tu X, Wang J, Huang AJ, Laskar MTR, et al. Utilizing BERT for information retrieval: survey, applications, resources, and challenges. ACM Comput Surv. 2024;56(7):1–33. doi:10.1145/3648471.
5. Abdel-Nabi H, Awajan A, Ali MZ. Deep learning-based question answering: a survey. Knowl Inf Syst. 2022;65(4):1399–485. doi:10.1007/s10115-022-01783-5.
6. Lin JC, Shao Y, Zhou Y, Pirouz M, Chen HC. A Bi-LSTM mention hypergraph model with encoding schema for mention extraction. Eng Appl Artif Intell. 2019;85:175–81. doi:10.1016/j.engappai.2019.06.005.
7. Raj M, Mishra N. Exploring new approaches for information retrieval through natural language processing. arXiv:2505.02199. 2025. doi:10.48550/arXiv.2505.02199.
8. Wang P, Gu J. Named entity recognition of electronic medical records based on BERT-BiLSTM-biaffine model. J Phys Conf Ser. 2023;2560(1):012044. doi:10.1088/1742-6596/2560/1/012044.
9. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805. 2019. doi:10.48550/arXiv.1810.04805.
10. Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R. ALBERT: a lite BERT for self-supervised learning of language representations. arXiv:1909.11942. 2020. doi:10.48550/arXiv.1909.11942.
11. Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT: a distilled version of BERT. arXiv:1910.01108. 2019. doi:10.48550/arXiv.1910.01108.
12. Clark K, Luong MT, Le QV, Manning CD. ELECTRA: pre-training text encoders as discriminators rather than generators. arXiv:2003.10555. 2020. doi:10.48550/arXiv.2003.10555.
13. Tho BD, Nguyen MT, Le DT, Ying LL, Inoue S, Nguyen TT. Improving biomedical Named Entity Recognition with additional external contexts. J Biomed Inform. 2024;156:104674. doi:10.1016/j.jbi.2024.104674.
14. Seow WL, Chaturvedi I, Hogarth A, Mao R, Cambria E. A review of named entity recognition: from learning methods to modelling paradigms and tasks. Artif Intell Rev. 2025;58(10):315. doi:10.1007/s10462-025-11321-8.
15. Nagpal A, Dasgupta R, Ganesan B. Fine grained classification of personal data entities with language models. In: Proceedings of the 5th Joint International Conference on Data Science & Management of Data (9th ACM IKDD CODS and 27th COMAD); 2022 Jan 8–10; Bangalore, India. p. 130–4. doi:10.1145/3493700.3493707.
16. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. arXiv:1706.03762. 2023. doi:10.48550/arXiv.1706.03762.
17. Tafjord O, Clark P. General-purpose question-answering with macaw. arXiv:2109.02593. 2021. doi:10.48550/arXiv.2109.02593.

18. Ghojogh B, Ghodsi A. Attention mechanism, transformers, BERT, and GPT: tutorial and survey. Preprint. 2020. doi:10.31219/osf.io/m6gcn.
19. Floridi L, Chiriatti M. GPT-3: its nature, scope, limits, and consequences. Mines Mach. 2020;30(4):681–94. doi:10.1007/s11023-020-09548-1.
20. Alomari A, Al-Shamayleh AS, Idris N, Qalid Md Sabri A, Alsmadi I, Omary D. Warm-starting for improving the novelty of abstractive summarization. IEEE Access. 2023;11:112483–501. doi:10.1109/access.2023.3322226.
21. Parikh S, Vohra Q, Tiwari M. Automated utterance generation. Proc AAAI Conf Artif Intell. 2020;34(8):13344–9. doi:10.1609/aaai.v34i08.7047.
22. Seo H, Kim E, Park M. Layer-wise semantic role labeling with the KR-BERT language model. Korean J Linguistics. 2022;47(3):445–66. [cited 2026 Jan 1]. Available from: https://www.dbpia.co.kr/Journal/articleDetail?nodeId=NODE11141876.
23. Trojovský P, Dehghani M. Walrus optimization algorithm: a new bio-inspired metaheuristic algorithm. Preprint. 2022. doi:10.21203/rs.3.rs-2174098/v1.
24. Shi P, Lin J. Simple BERT models for relation extraction and semantic role labeling. arXiv:1904.05255. 2019. doi:10.48550/arXiv.1904.05255.
25. Adnan K, Akbar R. An analytical study of information extraction from unstructured and multidimensional big data. J Big Data. 2019;6(1):91. doi:10.1186/s40537-019-0254-8.
26. Lilian JF, Sundarakantham K, Shalinie SM. QeCSO: design of hybrid Cuckoo Search based Query expansion model for efficient information retrieval. Sādhanā. 2021;46(3):181. doi:10.1007/s12046-021-01706-0.
27. Butt S, Ashraf N, Fahim H, Sidorov G, Gelbukh A. Transformer-based extractive social media question answering on TweetQA. CyS. 2021;25(1):23–32. doi:10.13053/cys-25-1-3897.
28. Rajpurkar P, Jia R, Liang P. Know what you don't know: unanswerable questions for SQuAD 2.0. In: Proceedings of the Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL). Melbourne, Australia: Association for Computational Linguistics; 2018. p. 784–9.
29. Bajaj P, Campos D, Craswell N, Deng L, Gao J, Liu X, et al. MS MARCO: a human-generated machine reading comprehension dataset. arXiv:1611.09268. 2016.
30. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: a robustly optimized BERT pretraining approach. arXiv:1907.11692. 2019.