Advancing Question Answering on Handwritten Documents

Aniket Pal[0000−0002−9971−8674], Ajoy Mondal[0000−0002−4808−8860], and C. V. Jawahar[0000−0001−6767−7057]

CVIT, IIIT, Hyderabad, India
aniket.pal@research.iiit.ac.in, (ajoy.mondal and jawahar)@iiit.ac.in

Abstract. Question-answering (QA) on handwritten documents is challenging but has valuable real-world applications. This paper presents a novel recognition-based QA approach that significantly improves accuracy over previous methods on handwritten datasets, including HW-SQuAD and BenthamQA. Our method integrates transformer-based retrieval and ensemble techniques, achieving Exact Match scores of 82.02% for HW-SQuAD and 69.1% for BenthamQA, with F1 Score improvements of 13.28% and 3.16%, respectively. It surpasses the previous best methods by 10.89% and 3.0%. Additionally, the document retrieval accuracy increased from 90.0% to 95.30% for HW-SQuAD and from 98.5% to 100% for BenthamQA. These results highlight the effectiveness of our approach in enhancing QA for handwritten documents. The data, model, and code are publicly available at https://github.com/phc2017002/improved_hw_squad.

Keywords: Handwritten documents · Question-Answering · Recognition-based · Ensemble · Document retrieval · BERT

1 Introduction

Question Answering (QA) [18,17] is a key NLP task that aims to provide correct answers to questions in natural language. Its applications include information retrieval, knowledge management, and intelligent personal assistants. The Stanford Question Answering Dataset (SQuAD) [18] is a widely-used benchmark for evaluating QA systems, consisting of Wikipedia articles and crowd-sourced questions, with answers as text spans from the articles. Answering questions on handwritten document images presents unique challenges compared to traditional text-based QA. Handwritten documents have complex layouts, varying writing styles, and potential noise and distortions, making information extraction difficult. Additionally, recognizing handwritten text is challenging due to its variability and ambiguity, requiring robust models to handle these issues effectively.

To address these challenges, the HW-SQuAD dataset [13] extends SQuAD to the handwritten domain. It includes synthetic handwritten document images paired with questions and answers, with answers spanning text within the documents. This dataset has driven research into QA systems capable of handling complexities of handwritten documents. The main approaches for HW-SQuAD are recognition-based methods, which convert handwritten images into machine-readable text for traditional QA techniques, and recognition-free methods, which directly process handwritten images using visual features and spatial layout information to find answer spans.

Previous works on HW-SQuAD have explored both recognition-based and recognition-free approaches. Minesh *et al.* [13] proposed a recognition-based method combining handwritten text recognition with a pre-trained language model. In the recognition-free domain, the same author introduced a visual QA model operating directly on handwritten document images. Despite these advancements, significant room for improvement remains in QA performance on handwritten documents. The previous recognition-based model consists of two main components: a document retriever using the naive TF-IDF (Term Frequency-Inverse Document Frequency) [19] algorithm to rank and select relevant documents based on the question, and a document reader utilizing the BERT (Bidirectional Encoder Representations from Transformers) QA [2] model to extract the answer span from the retrieved documents.

In this work, we propose a new *Document Retriever* that integrates TF-IDF with Sentence Transformer [20], leading to notable improvements in retrieval performance. We also apply advanced pre-processing techniques to enhance accuracy further. For the *Document Reader*, we use an ensemble approach combining the BERT QA model with SpanBERT [7] and DeBERTa [4] to achieve more accurate and robust answer extraction.

The main contributions of our work are as follows:

- We propose a novel document retriever that effectively combines NLP preprocessing, TF-IDF algorithm and Sentence Transformers, significantly improving the retrieval performance on the HW-SQuAD dataset.
- We employ an ensemble approach for the document reader component, leveraging the strengths of multiple extractive QA models, including BERT, SpanBERT, and DeBERTa, to achieve more robust and accurate answer extraction.
- Through extensive experiments, we demonstrate the superiority of our proposed approach, surpassing the previous state-of-the-art performance on the HW-SQuAD and BenthamQA datasets.

2 Related Works

The natural language processing (NLP) and information retrieval (IR) communities have been actively researching machine reading comprehension (MRC) and open domain question answering (QA). Large-scale datasets like SQuAD [18], MS MACRO [15], and Natural Questions [12] have driven the development of deep learning-based QA/MRC systems [2,5,22] capable of answering questions about given text corpora. Building on these advancements, our work focuses on answering questions on handwritten document images using recognition-based approaches.

Visual Question Answering (VQA) has recently gained significant attention in the computer vision community [1,22,3,10]. Early VQA datasets and methods often ignored text in images, focusing on visual aspects like objects and attributes. However, Gurari *et al.* [3] highlighted the necessity of reading and interpreting image text to answer many questions posed by visually impaired individuals, leading to the creation of datasets like Scene Text VQA [1] and TextVQA [10]. Our work differs in two main aspects — (i) it focuses on handwritten document images instead of "in the wild" images with scattered text tokens, and (ii) it aims to answer questions on a collection of document images rather than a single image.

Other relevant VQA works include VQA on charts and plots [8,9] and Textbook Question Answering (TQA) [10]. TQA answers questions based on text, diagrams, and images, but the text is computer-readable. For VQA on charts and plots, OCR is used to recognize text, but the text is sparse and in standard fonts, unlike the handwritten text in our case. Our work is inspired by the DocVQA dataset [14], which includes a variety of document images with printed, typewritten, handwritten, and digital text. While DocVQA Task 1 follows the standard VQA setting with textual answers, we propose an enhanced recognition-based approach for answering questions on handwritten document collections like HW-SQuAD and BenthamQA.

In information retrieval and keyword spotting, extensive efforts have been made in handwritten document indexing and retrieval [6,21]. The ImageCLEF 2016 Handwritten Scanned Document Retrieval challenge [21] focused on developing retrieval systems for handwritten documents. While similar to our work in using multi-token queries and retrieving document segments, this challenge differs in that queries are search terms, not natural language questions, and all query tokens must appear in the same order in the retrieved snippet. Kise *et al.* [11] addressed document retrieval for a QA system on printed document images using machine-printed English text, which is easier to recognize than handwritten text. Our work advances this by proposing an improved recognition-based approach for QA on handwritten document collections like HW-SQuAD and BenthamQA. DocVQA Task 2 [14] also deals with QA over a document collection but uses US candidate registration forms with the same template. Our focus is on handwritten documents with diverse content, aiming to return precise answer snippets rather than retrieving all documents required to answer the question correctly.

We propose an enhanced recognition-based approach for Question Answering on handwritten document collections, building on the method in [13]. We improved the *Document Retriever* with advanced pre-processing, TF-IDF, and Transformer models. Additionally, we enhanced the *Document Reader* using an Ensemble method. Our end-to-end pipeline achieves state-of-the-art performance on the HW-SQuAD and BenthamQA datasets.

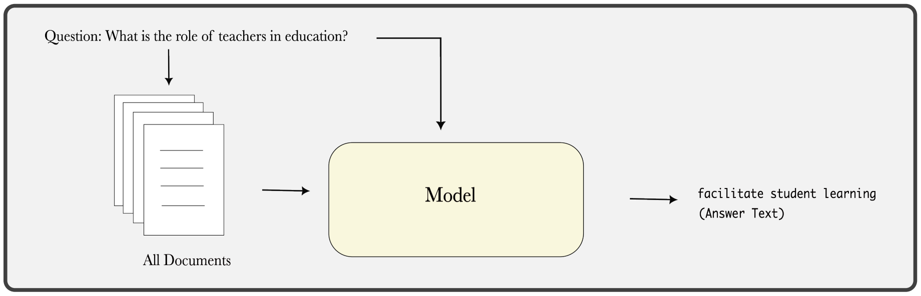

Question: What is the role of teachers in education?
All Documents
Model
facilitate student learning
(Answer Text)

Fig. 1: Visually illustrates an overview of our problem statement: The model receives a question and a collection of documents, and its task is to predict the correct answer.

3 Methodology

This section outlines our proposed end-to-end pipeline for question answering, illustrated in Fig. 1. The model receives a question and a collection of documents to predict the answer. For instance, when asked "What is the role of teachers in education?" along with all relevant documents, the model must predict the answer "facilitate student learning." The model can be either recognition-based or recognition-free. The existing recognition-based architecture [13] consists of two main components: (i) *Document Retriever* and (ii) *Document Reader*. The *Document Retriever* uses TF-IDF and pre-processing to select the top k documents from the entire collection. For the *Document Reader*, the BERT large model is employed.

We enhance the *Document Retriever* and *Document Reader* components in our proposed architecture. We incorporate NLP pre-processing techniques for document retrieval, including sentence transformers and TF-IDF. We use an ensemble of two large BERT models and one DeBERTa model for the *Document Reader*. These improvements significantly boost the performance of both document retrieval and reading. Fig. 2 illustrates the complete architecture of our proposed model.

3.1 Document Retriever Module

We introduce a novel technique that improves the traditional TF-IDF algorithm by integrating sentence transformers and NLP pre-processing steps.

NLP Pre-processing: It converts all characters to lowercase for consistency, keeps hyphens while removing other punctuation, retains only alphanumeric characters and hyphens, and then returns the cleaned and standardized text for improved analysis and NLP tasks.

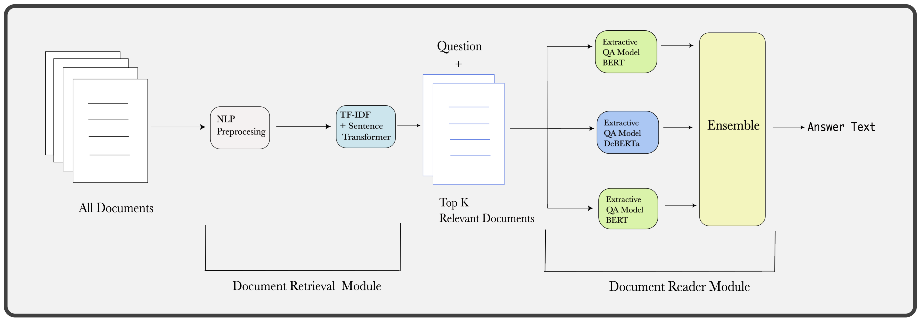


Fig. 2: Depicts the complete workflow and architecture of our proposed recognition-based approach. The *Document Retrieve* module comprises NLP pre-processing, TF-IDF, and Sentence Transformers. We implement an ensemble technique in the *Document Reader* module using two large BERT models and one DeBERTa large model.

Vectorization using TF-IDF: After pre-processing, we applied TF-IDF [19] for vectorization of text. Let V be the vocabulary of unique terms across all documents in a pre-processed document D_p. The TF-IDF vectorization can be represented as $\mathbf{M}_{\text{TF-IDF}} = \text{TF-IDF Vectorizer}(C)$, where $\mathbf{M}_{\text{TF-IDF}}$ is the resulting document term matrix and C is set of all pre-processed contexts. For each entry in $\mathbf{M}_{\text{TF-IDF}}$ matrix, we calculate the TF-IDF score and then vectorize them.

In the proposed *Document Retriever* module, we used TF-IDF vectorization combined with the Sentence Transformer encoding. We then applied cosine similarity to each method and merged them using a weighted approach to produce the final output.

Transformer Encoding: The Sentence Transformer model converts text documents into dense vector representations, forming a matrix $\mathbf{E}_C = \text{model}(C)$, where model is the Sentence Transformer and C represents the contexts. We used the all-MiniLM-L6-v2 (s) Sentence Transformer (a variant of MiniLM model) as it is small and efficient.

We compute cosine similarities between query vectors and context matrices from both TF-IDF and Sentence Transformer representations. Combining these similarities with predefined weights — 60% for TF-IDF and 40% for the Sentence Transformer — yields an ensemble similarity score. This approach integrates the strengths of both models, allowing us to retrieve the top N documents based on the highest ensemble scores. The entire *Document Retrieval* pipeline is given in Algorithm 1.

3.2 Document Reader Module

We propose an ensemble approach for the Document Reader module, as shown in Fig. 2. This ensemble combines the strengths of two large BERT models

Algorithm 1 Steps for Document Retrieval

```
1: Query Vectorization and Encoding
    • Input: Query $q$
    • Pre-process the query: $q' = \text{pre-process}(q)$
    • Vectorize using vectorizers: $\mathbf{q}_{\text{TF-IDF}} = \text{TF-IDF Vectorizer.transform}(q')$
    • Encode using transformer model: $\mathbf{e}_q = \text{ST.encode}(\text{model}, q')$
2: Cosine Similarity Calculation
    Compute cosine similarities between the query vectors and context matrices:
    • TF-IDF Cosine Similarity: $\mathbf{s}_{\text{TF-IDF}} = \cos(\mathbf{q}_{\text{TF-IDF}}, \mathbf{M}_{\text{TF-IDF}})$
    • Transformer Cosine Similarity: $\mathbf{s}_{\text{T}} = \cos(\mathbf{e}_q, \mathbf{E}_C)$
3: Ensemble Similarity Calculation
    • Combine similarity scores using predefined weights: $\mathbf{s}_{\text{E}} = 0.6 \cdot \mathbf{s}_{\text{TF-IDF}} + 0.4 \cdot \mathbf{s}_{\text{T}}$
4: Top-N Document Retrieval
    • Retrieve indices of top $n$ documents: $\text{top}_n = \text{argsort}(\mathbf{s}_{\text{E}})[-n :]$
    • Retrieve the corresponding contexts: $\text{top}_n\text{contexts} = [C_i \text{ for } i \text{ in top}_n]$
```

with different initializations and one large DeBERTa model. Both BERT and DeBERTa are transformer-based models excelling in natural language processing tasks, including question answering. Algorithm 2 details the workings of these models and their ensemble method. The algorithm has mainly two phases.

Training Phase: Each model (BERT1, BERT2, and DeBERTa) is fine-tuned separately on the QA task by generating hidden representations, computing start and end probabilities for answer spans, and updating model parameters to minimize the loss function.

Testing and Evaluation Phase: During inference, predictions for each question in the test and validation sets are generated by all three fine-tuned models. The ensemble prediction is the union of these answers. The algorithm evaluates performance by comparing predictions with ground truth answers and categorizing them as correct, similar, or incorrect. The HW-SQuAD dataset employs the mentioned training approach; however, in the case of BenthamQA, we solely evaluate the model. Additionally, we implemented an ensemble of three distinct architectures in the BenthamQA evaluation: the BERT-large, SpanBERT, and DeBERTa-large models.

This ensemble method enhances accuracy and reliability by combining multiple state-of-the-art models, leading to more robust results and capturing nuances individual models might miss.

4 Experimental setup

We evaluate our proposed model using the BenthamQA and HW-SQuAD datasets. Combining OCR texts, we created the context for our recognition-based pipeline with questions and answers derived from these datasets. After extracting context, questions, and answers, we implemented advanced pre-processing and our

Algorithm 2 Ensemble Approach for Document Reader Module

```
Initialize three models: Two BERT models (BERT1, BERT2) and one DeBERTa model
Training Phase:
for each model $m \in \{\text{BERT1}, \text{BERT2}, \text{DeBERTa}\}$ do
    • Fine-tune model $m$ on the question-answering task:
    • Generate hidden representations: $H_m = m(x)$, where $x = [x_1, x_2, \ldots, x_n]$ is the input sequence.
    • Compute start and end probabilities for each token: $\mathbf{P^{(m)}_{\text{start}}(i)} = \text{softmax}(\mathbf{W^{(m)}_{\text{start}}} \cdot \mathbf{h^{(m)}_i})$ and $\mathbf{P^{(m)}_{\text{end}}(i)} = \text{softmax}(\mathbf{W^{(m)}_{\text{end}}} \cdot \mathbf{h^{(m)}_i})$, where $W^{(m)}_{\text{start}}$ and $W^{(m)}_{\text{end}}$ are learnable weight matrices.
    • Update model parameters to minimize the loss function
end for
Testing and Evaluation Phase:
for each question in the dev and test set do
    • Generate predictions using each fine-tuned model:
        $\mathbf{A_1} = \text{BERT1}(\text{question}, \text{context})$
        $\mathbf{A_2} = \text{BERT2}(\text{question}, \text{context})$
        $\mathbf{A_3} = \text{DeBERTa}(\text{question}, \text{context})$
    • Compute ensemble prediction: $\mathbf{A}_{\text{ensemble}} = \mathbf{A_1} \cup \mathbf{A_2} \cup \mathbf{A_3}$
    • Evaluate ensemble performance:
    • Compare $A_{\text{ensemble}}$ with ground truth answer
    • Categorize prediction as:
    • Correct: Exact match with ground truth
    • Similar: Significant overlap with ground truth
    • Incorrect: No match or significant overlap
end for
Compute overall performance metrics (e.g., accuracy, F1 score)
```

proposed Document Retriever. We then converted the output to the SQuAD dataset format for the recognition-based model. The HW-SQuAD dataset contains over 84,000 QA pairs, while BenthamQA has 200 pairs. We used the same training, validation, and testing split as in [13], fine-tuning our model on HW-SQuAD. BenthamQA was used only for evaluation. The training was conducted using three 2080 ti and one 4080 Nvidia RTX graphics cards. We employed AdamW and Adam for BERT large and DeBERTa, respectively. We used the simple transformer library [16] for fine-tuning and evaluation, while the haystack library is used in [13].

4.1 Evaluation Metrics

We used top-5 accuracy for evaluating the performance of the Document Retriever (document retrieval task), as implemented in [13]. It is calculated as:

$$top - 5\ accuracy = \frac{1}{N}\sum_{i=1}^{N} I(ground\_truth_i \in top\_5\_predictions_i), \qquad (1)$$

where N is the total number of instances and I is the indicator function that equals one if the ground truth is present in the top 5 predictions, and 0 otherwise. We employed two evaluation metrics for Document Reader: F1 score [23] and Exact Match (EM) [18,13].

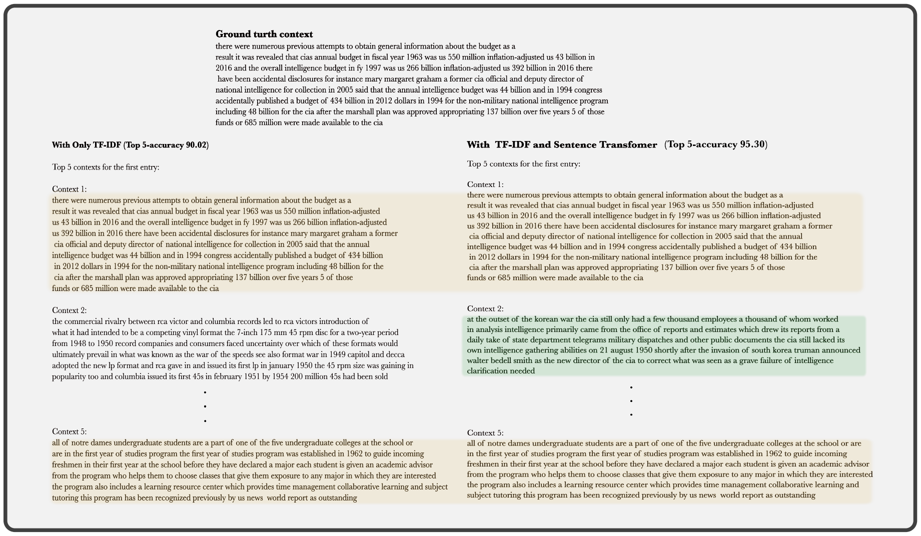
5 Results and Analysis

5.1 Document Retrievals by Document Retriever

Table 1: Show results of document retrieval and QA tasks on HW-SQuAD and BenthamQA datasets.

Method	**Document Retrieval Task**		**QA Task**			
	HW-SQuAD	**BenthamQA**	**HW-SQuAD**		**BenthamQA**	
	top-5 Acc.	**top-5 Acc.**	**F1**	**EM**	**F1**	**EM**
Method [13]	90.20	98.50	76.82	70.73	78.41	66.00
Our	**95.30**	**100.00**	**90.10**	**82.02**	**81.57**	**69.10**

Quantitative Results: Table 1 compares document retrieval and question-answering tasks on the HW-SQuAD and BenthamQA datasets. We examine the impact of incorporating sentence transformers and NLP pre-processing techniques alongside the TF-IDF algorithm for document retrieval. The existing technique [13] relied solely on the TF-IDF pre-processing approach, achieving a top-5 accuracy of 90.0%. Our proposed method improves document retrieval accuracy by including sentence transformers and NLP pre-processing techniques. On the HW-SQuAD dataset, our approach attains a top-5 accuracy of 95.3%, outperforming the previous model [13] by 5.1%. Similarly, on the BenthamQA dataset, our method achieves a top-5 accuracy of 100%. These results underscore the effectiveness of leveraging sentence transformers and NLP pre-processing in capturing semantic similarities between the question and the document, enabling more precise retrieval of relevant documents.



Ground turth context

there were numerous previous attempts to obtain general information about the budget as a result it was revealed that cias annual budget in fiscal year 1963 was us 550 million inflation-adjusted us 43 billion in 2016 and the overall intelligence budget in fy 1997 was us 266 billion inflation-adjusted us 392 billion in 2016 there have been accidental disclosures for instance mary margaret graham a former cia official and deputy director of national intelligence for collection in 2005 said that the annual intelligence budget was 44 billion and in 1994 congress accidentally published a budget of 434 billion in 2012 dollars in 1994 for the non-military national intelligence program including 48 billion for the cia after the marshall plan was approved appropriating 137 billion over five years 5 of those funds or 685 million were made available to the cia

With Only TF-IDF (Top 5-accuracy 90.02)

Top 5 contexts for the first entry:

Context 1:
there were numerous previous attempts to obtain general information about the budget as a result it was revealed that cias annual budget in fiscal year 1963 was us 550 million inflation-adjusted us 43 billion in 2016 and the overall intelligence budget in fy 1997 was us 266 billion inflation-adjusted us 392 billion in 2016 there have been accidental disclosures for instance mary margaret graham a former cia official and deputy director of national intelligence for collection in 2005 said that the annual intelligence budget was 44 billion and in 1994 congress accidentally published a budget of 434 billion in 2012 dollars in 1994 for the non-military national intelligence program including 48 billion for the cia after the marshall plan was approved appropriating 137 billion over five years 5 of those funds or 685 million were made available to the cia

Context 2:
the commercial rivalry between rca victor and columbia records led to rca victors introduction of what it had intended to be a competing vinyl format the 7-inch 175 mm 45 rpm disc for a two-year period from 1948 to 1950 record companies and consumers faced uncertainty over which of these formats would ultimately prevail in what was known as the war of the speeds see also format war in 1949 capitol and decca adopted the new lp format and rca gave in and issued its first lp in january 1950 the 45 rpm size was gaining in popularity too and columbia issued its first 45s in february 1951 by 1954 200 million 45s had been sold

⋮

Context 5:
all of notre dames undergraduate students are a part of one of the five undergraduate colleges at the school or are in the first year of studies program the first year of studies program was established in 1962 to guide incoming freshmen in their first year at the school before they have declared a major each student is given an academic advisor from the program who helps them to choose classes that give them exposure to any major in which they are interested the program also includes a learning resource center which provides time management collaborative learning and subject tutoring this program has been recognized previously by us news world report as outstanding

With TF-IDF and Sentence Transfomer (Top 5-accuracy 95.30)

Top 5 contexts for the first entry:

Context 1:
there were numerous previous attempts to obtain general information about the budget as a result it was revealed that cias annual budget in fiscal year 1963 was us 550 million inflation-adjusted us 43 billion in 2016 and the overall intelligence budget in fy 1997 was us 266 billion inflation-adjusted us 392 billion in 2016 there have been accidental disclosures for instance mary margaret graham a former cia official and deputy director of national intelligence for collection in 2005 said that the annual intelligence budget was 44 billion and in 1994 congress accidentally published a budget of 434 billion in 2012 dollars in 1994 for the non-military national intelligence program including 48 billion for the cia after the marshall plan was approved appropriating 137 billion over five years 5 of those funds or 685 million were made available to the cia

Context 2:
at the outset of the korean war the cia still only had a few thousand employees a thousand of whom worked in analysis intelligence primarily came from the office of reports and estimates which drew its reports from a daily take of state department telegrams military dispatches and other public documents the cia still lacked its own intelligence gathering abilities on 21 august 1950 shortly after the invasion of south korea truman announced walter bedell smith as the new director of the cia to correct what was seen as a grave failure of intelligence clarification needed

⋮

Context 5:
all of notre dames undergraduate students are a part of one of the five undergraduate colleges at the school or are in the first year of studies program the first year of studies program was established in 1962 to guide incoming freshmen in their first year at the school before they have declared a major each student is given an academic advisor from the program who helps them to choose classes that give them exposure to any major in which they are interested the program also includes a learning resource center which provides time management collaborative learning and subject tutoring this program has been recognized previously by us news world report as outstanding

Fig. 3: Compares the two approaches, method [13] and ours, for document retrieval tasks. The retrieved common paragraphs (from HW-SQuAD) for both these methods are highlighted with light brown. Improved retrieved paragraphs are highlighted with light green.

Qualitative Results: Fig. 3 visually illustrates the performance comparison between method [13] (using only TF-IDF) and our proposed approach (incorporating TF-IDF, sentence transformers, and NLP-based pre-processing). Both methods successfully retrieve the required paragraph (exact match with the query text) at the top rank. However, in the proposed method, the subsequent retrieved paragraphs exhibit higher relevance in terms of semantic and thematic consistency with the query text than the method [13].

Semantic Relevance and Thematic Consistency — The proposed approach demonstrates improved semantic relevance and thematic consistency in paragraph retrieval compared to the method [13]. It retrieves contexts more closely related to terms like "CIA" and "intelligence" agencies, maintaining better topical coherence. For instance, context 2 retrieved by the proposed approach is more semantically aligned with the query context. In contrast, the method [13], which relies solely on keyword matching, often retrieves less relevant or thematically inconsistent contexts.

Reduced Reliance on Exact Keyword Matching — In addition to improving semantic relevance and thematic consistency, the proposed approach enables the retrieval of relevant contexts that may use synonyms, paraphrases, or related terms instead of the exact keywords found in the query. By capturing the semantic similarity between words and phrases, this approach can identify pertinent contexts that would be overlooked by a purely keyword-based method like TF-IDF.

In conclusion, the proposed approach incorporating TF-IDF, sentence transformers, and NLP pre-processing offers significant advantages over using only TF-IDF. The proposed approach retrieves more relevant and coherent contexts by capturing semantic meaning and thematic consistency and reducing the reliance on exact keyword matching.

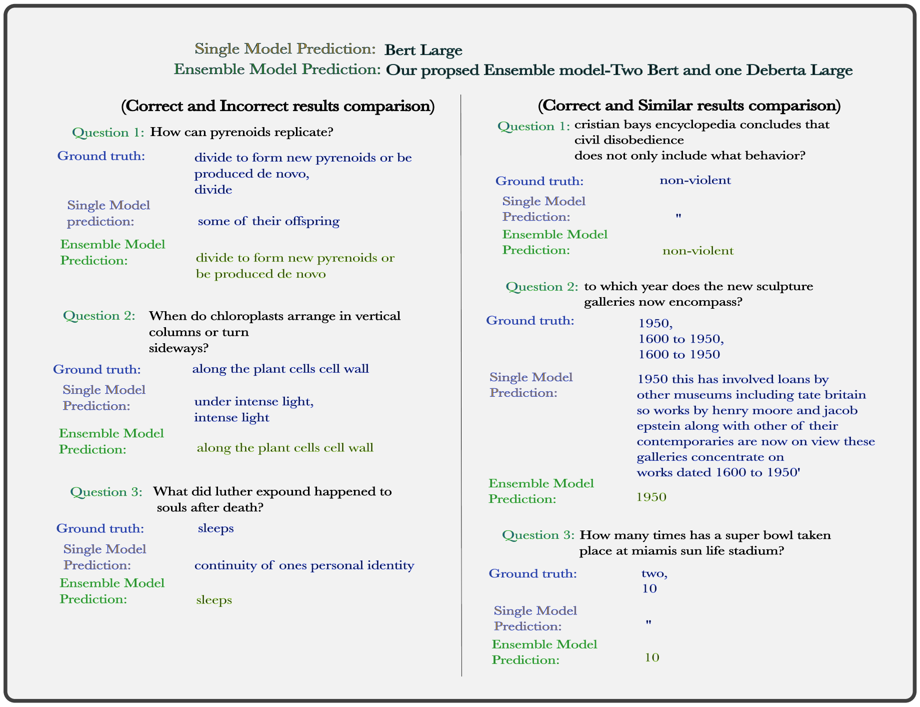
5.2 Question Answering by Ensemble Method on Document Reader

Quantitative Results: Table 1 also presents a performance comparison between the baseline (single) method [13] and the proposed (ensemble) method for question-answering tasks on both HW-SQuAD and BenthamQA datasets. The results show that the proposed model demonstrates superior performance across both datasets. On the HW-SQuAD dataset, it achieves an F1 score of 90.10% and an Exact Match (EM) of 82.02%, significantly improving over the baseline model's [13] 76.82% F1 and 70.73% EM. Similarly, on the BenthamQA dataset, our method attains an F1 score of 81.57% and an EM of 69.0%, substantially outperforming the baseline's 78.41% F1 and 66% EM.

The enhanced performance of the proposed method can be attributed to TF-IDF, sentence transformer, and NLP pre-processing in the document retriever module and an ensemble approach (combination of BERT and DeBERTa) in the document reader module, building upon the standard TF-IDF + BERT baseline for question answering tasks. The ensemble model's exceptional results, approaching 90.0% F1 and exceeding 80.0% Exact Match, establish a new benchmark for performance on these challenging datasets. The method introduced in this study has the potential to advance state-of-the-art question-answering systems for handwritten documents.

Qualitative Results: Figs. 4 and 5 show a few results obtained by baseline method [13] and the proposed method for HW-SQuAD and BenthamQA datasets, respectively, for the question-answering task. For the HW-SQuAD dataset, the proposed model consistently predicts correct answers, while the baseline method [13] predicts erroneous answers for most cases. For example, where the baseline model inaccurately predicts "some of their offspring," the proposed method accurately identifies "divide to form new pyrenoids or be produced de novo". In contrast, the proposed model accurately matches the ground truth. This precision is also evident in questions requiring detailed understanding, such as correctly predicting "along the plant cells cell wall" instead of "under intense light". Additionally, for broader categorical answers like "humid subtropical," the proposed model refines the baseline model's vague prediction of "warm" to the precise ground truth. This meticulous correction and alignment with the expected answers highlight the proposed model's enhanced interpretative capabilities.

The effectiveness of the proposed model is equally evident in the BenthamQA dataset. It corrects broader and less accurate predictions of the baseline model and ensures precision in complex queries. For instance, it refines the single model's generic prediction of "Offenses against Property Theft" to the specific



Single Model Prediction: Bert Large
Ensemble Model Prediction: Our propsed Ensemble model-Two Bert and one Deberta Large

(Correct and Incorrect results comparison)

Question 1: How can pyrenoids replicate?

Ground truth:	divide to form new pyrenoids or be produced de novo, divide
Single Model prediction:	some of their offspring
Ensemble Model Prediction:	divide to form new pyrenoids or be produced de novo

Question 2: When do chloroplasts arrange in vertical columns or turn sideways?

Ground truth:	along the plant cells cell wall
Single Model Prediction:	under intense light, intense light
Ensemble Model Prediction:	along the plant cells cell wall

Question 3: What did luther expound happened to souls after death?

Ground truth:	sleeps
Single Model Prediction:	continuity of ones personal identity
Ensemble Model Prediction:	sleeps

(Correct and Similar results comparison)

Question 1: cristian bays encyclopedia concludes that civil disobedience does not only include what behavior?

Ground truth:	non-violent
Single Model Prediction:	"
Ensemble Model Prediction:	non-violent

Question 2: to which year does the new sculpture galleries now encompass?

Ground truth:	1950, 1600 to 1950, 1600 to 1950
Single Model Prediction:	1950 this has involved loans by other museums including tate britain so works by henry moore and jacob epstein along with other of their contemporaries are now on view these galleries concentrate on works dated 1600 to 1950'
Ensemble Model Prediction:	1950

Question 3: How many times has a super bowl taken place at miamis sun life stadium?

Ground truth:	two, 10
Single Model Prediction:	"
Ensemble Model Prediction:	10

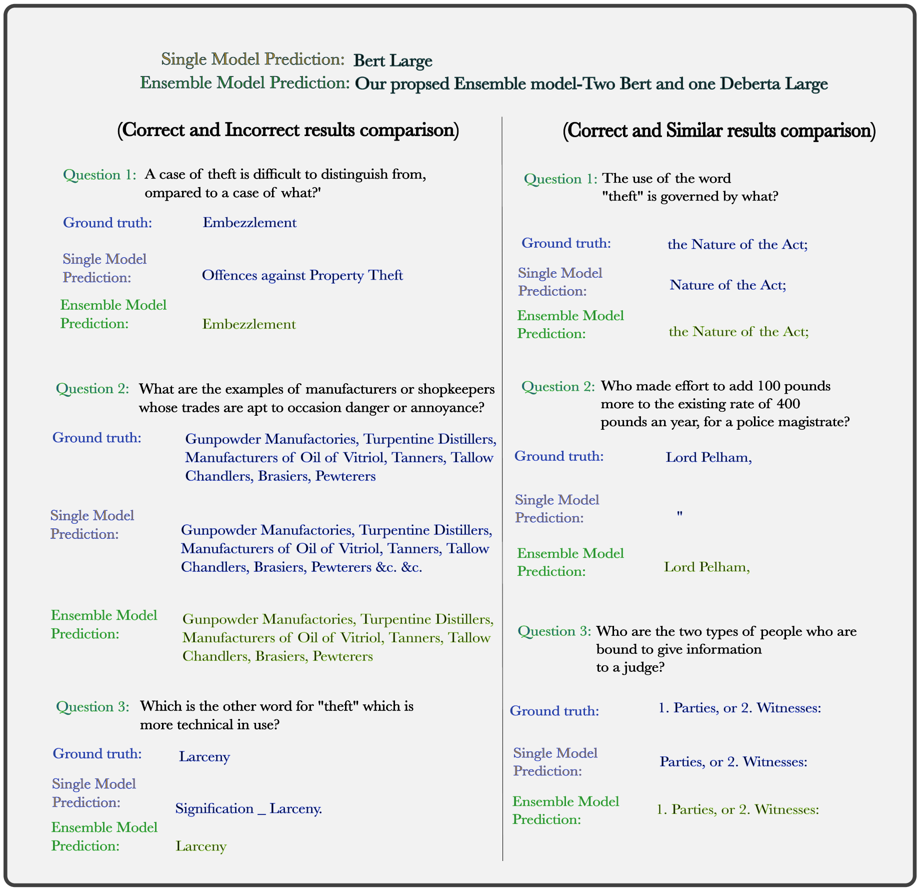
Fig. 4: Shows the comparison between the predicted results of the baseline (single) method and the proposed (ensemble) method for HW-SQuAD dataset. The left side presents the results obtained by the single model, while the right side displays the results from the ensemble method.

term "Embezzlement", demonstrating its ability to grasp nuanced legal terminology. Additionally, for questions requiring multiple valid responses, such as listing various historical manufacturers or recognizing titles like "Lord Pelham", the ensemble model accurately captures all relevant details, showcasing its comprehensive understanding and reliability.

Overall, the proposed model's ability to minimize incorrect and similar erroneous results significantly enhances its reliability and precision. Combining multiple models' strengths ensures predictions are more accurate and consistent with the ground truth. This integration enables the ensemble model to capture broader linguistic and contextual nuances, ultimately leading to more robust and dependable AI systems.

5.3 Ablation Studies

Quantitative Results: Table 2 illustrates the effectiveness of multiple components in the proposed Document retriever and Reader modules over baseline method [13]. In Table 2(a), we examined the impact of NLP based pre-processing



Single Model Prediction: Bert Large
Ensemble Model Prediction: Our propsed Ensemble model-Two Bert and one Deberta Large

(Correct and Incorrect results comparison)

Question 1: A case of theft is difficult to distinguish from, ompared to a case of what?'

Ground truth:	Embezzlement
Single Model Prediction:	Offences against Property Theft
Ensemble Model Prediction:	Embezzlement

Question 2: What are the examples of manufacturers or shopkeepers whose trades are apt to occasion danger or annoyance?

Ground truth:	Gunpowder Manufactories, Turpentine Distillers, Manufacturers of Oil of Vitriol, Tanners, Tallow Chandlers, Brasiers, Pewterers
Single Model Prediction:	Gunpowder Manufactories, Turpentine Distillers, Manufacturers of Oil of Vitriol, Tanners, Tallow Chandlers, Brasiers, Pewterers &c. &c.
Ensemble Model Prediction:	Gunpowder Manufactories, Turpentine Distillers, Manufacturers of Oil of Vitriol, Tanners, Tallow Chandlers, Brasiers, Pewterers

Question 3: Which is the other word for "theft" which is more technical in use?

Ground truth:	Larceny
Single Model Prediction:	Signification _ Larceny.
Ensemble Model Prediction:	Larceny

(Correct and Similar results comparison)

Question 1: The use of the word "theft" is governed by what?

Ground truth:	the Nature of the Act;
Single Model Prediction:	Nature of the Act;
Ensemble Model Prediction:	the Nature of the Act;

Question 2: Who made effort to add 100 pounds more to the existing rate of 400 pounds an year, for a police magistrate?

Ground truth:	Lord Pelham,
Single Model Prediction:	"
Ensemble Model Prediction:	Lord Pelham,

Question 3: Who are the two types of people who are bound to give information to a judge?

Ground truth:	1. Parties, or 2. Witnesses:
Single Model Prediction:	Parties, or 2. Witnesses:
Ensemble Model Prediction:	1. Parties, or 2. Witnesses:

Fig. 5: Shows the comparison between the predicted results of the baseline (single) method and the proposed (ensemble) method for BenthamQA dataset. The left side presents the results obtained by the single model, while the right side displays the results from the ensemble method.

and Sentence Transformer on our Document Retriever Module. The baseline model using only TF-IDF and basic pre-processing achieved a top-5 accuracy of 90.2%. Implementing NLP pre-processing techniques increased the accuracy to 90.8%. Further incorporating the Sentence Transformer resulted in a 5% performance boost. For the BenthamQA dataset, NLP pre-processing improved top-5 accuracy by approximately 0.5%, and adding the Sentence Transformer achieved a perfect 100% accuracy.

Table 2(b) highlights the effectiveness of multiple components in our proposed Document Reader module. In the case of BenthamQA, we initially applied a single BERT model, achieving a 63.21% EM and 77.46% F1 score without any pre-processing. Adding our proposed pre-processing to the retrieval pipeline im-

Table 2: Effect of multiple components in our Document Retriever and Reader modules. ST indicates Sentence Transformer, pp. indicates pre-processing, and En. indicates Ensemble.

(a) Effect of multiple components in the proposed Document Retriever module.

Components	top-5 Acc.	
	HW-SQuAD	BenthamQA
TF-IDF (baseline)	90.20	98.50
TF-IDF+pp.	90.80	98.94
TF-IDF+pp.+ST	95.30	100.00

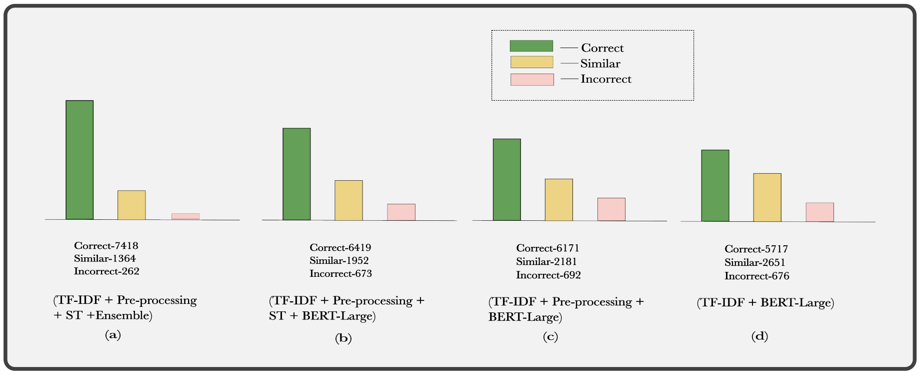
(b) Effect of multiple components in the proposed Document Reader module.

Components	HW-SQuAD		BenthamQA	
	F1	EM	F1	EM
TF-IDF+BERT	76.82	70.20	78.41	66.00
TF-IDF+BERT*	77.46	63.21	64.75	53.00
TF-IDF+pp.+BERT	81.18	68.35	73.96	59.20
TF-IDF+pp.+ST+BERT	83.20	71.33	73.96	59.20
TF-IDF+pp.+ST+En.	90.10	82.02	81.57	69.10

proved the scores to 68.21% EM and 77.46% F1. Incorporating the Sentence Transformer (ST) further enhanced performance, reaching an 71.33% EM, and 83.20% F1 score demonstrating the effectiveness of semantic information for retrieving more relevant documents. Similarly, for HW-SQuAD, Our baseline implementation achieved 53% EM and 64.75% F1 scores. Adding pre-processing and the Sentence Transformer to the Document Retriever, and fine-tuning BERT large, improved the scores to 59.2% EM and 73.96% F1. Finally, the ensemble approach combines three extractive QA models (two BERT and one DeBERTa large) with retrieval enhancements. The TF-IDF + ST + Ensemble model achieved an impressive 90.10% F1 score and 82.02% EM, significantly outperforming the baseline for HW-SQuAD. On the BenthamQA, the ensemble method reached over 69.0% EM and 81.57% F1. These results underscore the importance of improved document retrieval and the ensemble strategy in the proposed approach.

Qualitative Results: Fig. 6 visually presents the effectiveness of the multiple components in our proposed method for document retrieval tasks. Our proposed model yields 7,418 correct matches, 1,364 similar matches, and 262 incorrect matches, representing the lowest number of identical and incorrect matches among all evaluated models. The baseline model produces 676 incorrect matches, nearly three times greater than the proposed method. Adding a Sentence Transformer to the Document Retriever stage increases correct matches to 6,419, an improvement of over 700 compared to the baseline model, and reduces incorrect matches to 1,952. An ensemble technique further boosts correct matches to 7,419 and reduces incorrect matches to 1,364.

Adding pre-processing and Sentence Transformer to the Document Retriever significantly improves the quality of retrieved documents. Combined with a single BERT large model, this enhancement notably boosts the Document Reader's performance. Implementing an ensemble method further enhances semantic understanding, reducing incorrect matches by three-fold and similar matches by half compared to the baseline model.



(TF-IDF + Pre-processing + ST +Ensemble) (a)

(TF-IDF + Pre-processing + ST + BERT-Large) (b)

(TF-IDF + Pre-processing + BERT-Large) (c)

(TF-IDF + BERT-Large) (d)

Fig. 6: Visually illustrates the improved performance of the proposed method over the baseline model for document retrieval tasks. It increases the number of correct and reduces the similar and incorrect retrievals.

The ablation studies confirm that combining semantic similarity-enhanced retrieval with an ensemble of strong reader models achieves state-of-the-art performance on the HW-SQuAD benchmark.

6 Conclusions

This paper introduces a novel approach for answering questions on handwritten documents by integrating advanced document retrieval techniques with an ensemble of extractive QA models. Our enhanced retriever, combining TF-IDF and sentence transformers, significantly boosts retrieval performance. The ensemble-based reader, using BERT and DeBERTa large models, ensures accurate and robust answer extraction. The experimental results on HW-SQuAD and BenthamQA datasets show that our approach outperforms the baseline recognition-based method, setting a new benchmark for QA performance on handwritten documents. However, there is still potential for further improvement, including exploring recognition-free methods and addressing challenges like noise and diverse writing styles. Our work represents a major advancement in question-answering for handwritten documents. By introducing a novel approach and achieving superior results, we have advanced the state of the art and paved the way for further research in this challenging and crucial area of natural language processing.

References

1. Biten, A.F., Tito, R., Mafla, A., Gomez, L., Rusiñol, M., Jawahar, C., Valveny, E., Karatzas, D.: Scene text visual question answering. In: ICCV (2019)
2. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT (2019)

3. Gurari, D., Li, Q., Stangl, A.J., Guo, A., Lin, C., Grauman, K., Luo, J., Bigham, J.P.: Vizwiz grand challenge: Answering visual questions from blind people. In: CVPR (2018)
4. He, P., Liu, X., Gao, J., Chen, W.: Deberta: Decoding-enhanced bert with disentangled attention. In: ICLR (2021)
5. Hermann, K.M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., Blunsom, P.: In: NeurIPS (2015)
6. Jain, A., Namboodiri, A.: Indexing and retrieval of on-line handwritten documents. In: ICDAR (2003)
7. Joshi, M., Chen, D., Liu, Y., Weld, D.S., Zettlemoyer, L., Levy, O.: SpanBERT: Improving pre-training by representing and predicting spans. TACL (2020)
8. Kafle, K., Price, B., Cohen, S., Kanan, C.: Dvqa: Understanding data visualizations via question answering. In: CVPR (2018)
9. Kahou, S.E., Michalski, V., Atkinson, A., Kádár, Á., Trischler, A., Bengio, Y.: Figureqa: An annotated figure dataset for visual reasoning. In: ICLR (2018)
10. Kembhavi, A., Seo, M., Schwenk, D., Choi, J., Farhadi, A., Hajishirzi, H.: Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In: CVPR (2017)
11. Kise, K., Fukushima, S., Matsumoto, K.: Document image retrieval for qa systems based on the density distributions of successive terms. IEICE Trans. on inf. and sys. **88-D**, 1843–1851 (2005)
12. Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., Toutanova, K., Jones, L., Kelcey, M., Chang, M.W., Dai, A.M., Uszkoreit, J., Le, Q., Petrov, S.: Natural questions: A benchmark for question answering research. ACL (2019)
13. Mathew, M., Gomez, L., Karatzas, D., Jawahar, C.V.: Asking questions on handwritten document collections. IJDAR **24**(3), 235–249 (2021)
14. Mathew, M., Karatzas, D., Jawahar, C.V.: Docvqa: A dataset for vqa on document images. In: WACV (2021)
15. Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., Deng, L.: MS MARCO: A human generated machine reading comprehension dataset. In: NIPS workshop (2016)
16. Rajapakse, T.C.: Simple transformers. `https://github.com/ThilinaRajapakse/simpletransformers` (2019)
17. Rajpurkar, P., Jia, R., Liang, P.: Know what you don't know: Unanswerable questions for SQuAD. In: Gurevych, I., Miyao, Y. (eds.) ACL (2018)
18. Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: SQuAD: 100,000+ questions for machine comprehension of text. In: EMNLP (2016)
19. Ramos, J.E.: Using tf-idf to determine word relevance in document queries. In: Proceedings of ICML (2003)
20. Reimers, N., Gurevych, I.: Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In: EMNLP-IJCNLP (2019)
21. Villegas, M., Puigcerver, J., Toselli, A.H., Sánchez, J., Vidal, E.: Overview of the imageclef 2016 handwritten scanned document retrieval task. In: Working Notes of CLEF (2016)
22. Wang, S., Jiang, J.: Machine comprehension using match-LSTM and answer pointer. In: ICLR (2017)
23. Wikipedia contributors: F-score — Wikipedia, the free encyclopedia (2024), `https://en.wikipedia.org/w/index.php?title=F-score&oldid=1251544928`, [Online; accessed 26-October-2024]